

CLUSTERING, CODING, AND THE CONCEPT OF SIMILARITY

L. THORNE MCCARTY
DEPARTMENT OF COMPUTER SCIENCE
RUTGERS UNIVERSITY

ABSTRACT: This paper develops a theory of *clustering* and *coding* which combines a geometric model with a probabilistic model in a principled way. The geometric model is a Riemannian manifold with a Riemannian metric, $g_{ij}(\mathbf{x})$, which we interpret as a measure of *dissimilarity*. The probabilistic model consists of a stochastic process with an invariant probability measure which matches the density of the sample input data. The link between the two models is a potential function, $U(\mathbf{x})$, and its gradient, $\nabla U(\mathbf{x})$. We use the gradient to define the dissimilarity metric, which guarantees that our measure of dissimilarity will depend on the probability measure. Finally, we use the dissimilarity metric to define a coordinate system on the embedded Riemannian manifold, which gives us a low-dimensional encoding of our original data.

1. INTRODUCTION.

Clustering algorithms have been studied for several decades [DH73], and they remain one of the main ingredients in unsupervised learning [DHS01]. Intuitively, a *cluster* is both a geometric concept (e.g., a low-dimensional region in a high-dimensional space) and a probabilistic concept (e.g., a region of the input space in which the sample data density is high).

Recently, a variant of the traditional clustering algorithms has attracted some attention, under the rubric of *manifold learning*: [TSL00] [RS00] [BN03]. In this variant, the learning task is to construct a low-dimensional *manifold*, embedded in the original high-dimensional space, on which the probability density of the input data is high. For

example, in one recent paper, Rifai, et al. [RDV⁺11], outline three hypotheses that motivate much of this work:

1. The **semi-supervised learning hypothesis**, according to which learning aspects of the input distribution $p(x)$ can improve models of the conditional distribution of the supervised target $p(y|x)$, i.e., $p(x)$ and $p(y|x)$ share something ... [citations omitted]
2. The **(unsupervised) manifold hypothesis**, according to which real world data presented in high dimensional spaces is likely to concentrate in the vicinity of non-linear sub-manifolds of much lower dimensionality ... [citations omitted]
3. The **manifold hypothesis for classification**, according to which points of different classes are likely to concentrate along different sub-manifolds, separated by low density regions of the input space.

The authors then present a “Contractive Auto-Encoder (CAE)” algorithm to exploit these hypotheses, and they combine this with an existing supervised learning algorithm to produce what they call a “Manifold Tangent Classifier (MTC),” which performs very well on several datasets. It is interesting to note that these algorithms are based, explicitly, on concepts from differential geometry, but they draw only implicitly on probability theory. The informal language of probability theory abounds. For example, the authors write that the “data density concentrates near low-dimensional manifolds” and “different classes correspond to disjoint manifolds separated by low density” (see abstract). But there is no explicit probability model in the paper.

In this paper, we will develop a theory of *clustering* and *coding* which combines a geometric model with a probabilistic model in a principled way. The geometric model is a Riemannian manifold with a Riemannian metric, $g_{ij}(\mathbf{x})$, which we interpret as a measure of *dissimilarity*. The probabilistic model consists of a stochastic process with an invariant probability measure which matches the density of the sample input data. The link between the two models is a potential function, $U(\mathbf{x})$, and its gradient, $\nabla U(\mathbf{x})$. We use the gradient to define the dissimilarity metric, which guarantees that our measure of dissimilarity will depend on the probability measure. Roughly speaking, the dissimilarity will be small in a region in which the probability density is high, and vice versa. Finally, we use the dissimilarity metric to define a coordinate system on the embedded Riemannian manifold, which gives us a low-dimensional encoding of our original data.

Section 2 reviews the “Mathematical Background” of the paper, including several theorems which will play a central role in the subsequent discussion. Section 3 then discusses “Prototype Coding,” our overall model, and explains how the dissimilarity metric and the low-dimensional coordinate system are related to the stochastic process with an invariant probability measure. Section 4 investigates the differential geometry component of the model more carefully, with a focus on the important concept of an “Integral Manifold.” At this point in the paper, we restrict our analysis to \mathbf{R}^3 rather than \mathbf{R}^n , although we will see later (in Section 7) that this is not actually a limitation on the scope of the theory. Instead, the restriction to three dimensions simplifies our calculations, and makes them much easier to visualize. Accordingly, in Section 5, we present the results of a number of experiments using *Mathematica*, including some full-color three-dimensional graphics of several examples which are intended to aid our intuitions about the main elements of the theory. Section 6 discusses an interesting technical result, which also helps to link the geometric model to the probabilistic model. Finally, Section 7 discusses “Future Work,” including a further analysis of the connections between the present theory and the current literature on manifold learning.

2. MATHEMATICAL BACKGROUND.

Let’s start with a model that will be familiar to most physicists: the Feynman-Kac formula [Fey48] [Kac49]. We will write this formula as follows:

$$(1) \quad u(t, \mathbf{x}) = \int_{\Omega} f(X_t) \exp \left[- \int_0^t V(X_s) ds \right] \mathcal{W}_{\mathbf{x}}(dX)$$

Here, $X_t \equiv X(t, \omega)$ denotes a continuous path in \mathbf{R}^n , and $\mathcal{W}_{\mathbf{x}}$ denotes Wiener measure over all such paths beginning at $X_0 = \mathbf{x}$. If $V: \mathbf{R}^n \rightarrow \mathbf{R}$ is bounded below, then $u(t, \mathbf{x})$ is a solution to the Cauchy initial value problem:

$$(2) \quad \frac{\partial u}{\partial t} = \frac{1}{2} \Delta u - V(\mathbf{x}) u \quad \text{with} \quad u(0, \cdot) = f$$

in which Δ denotes the standard Laplacian in Cartesian coordinates. Conversely, any bounded solution to (2) is equal to the function defined by (1). See [Str93], Section 4.3. Now, following Feynman’s heuristic picture of formula (1), we can write a discrete approximation to Wiener measure as:

$$\int \exp \left[- \sum_{k=1}^m \frac{s_k - s_{k-1}}{2} \left(\frac{|X(s_k) - X(s_{k-1})|}{s_k - s_{k-1}} \right)^2 \right] dX(s_1) \dots dX(s_m),$$

multiplied by a normalization factor, so that the exponential function in the integrand of (1) could be viewed, in the limit, as:

$$(3) \quad \exp \left[- \int_0^t \frac{1}{2} |\dot{X}(s)|^2 + V(X(s)) \, ds \right]$$

See [Str93], Section 4.2, or [Str11], Section 8.1. The quantity inside the integral sign is, of course, the Hamiltonian of a classical dynamical system with the potential function: $V(\mathbf{x})$.

This model obviously possesses some of the properties that we want: Equations (1) and (2) specify a stochastic process that depends on the potential function, $V(\mathbf{x})$, and the exponent in formula (3) can be interpreted as an expression in differential geometry, which also depends on $V(\mathbf{x})$. Furthermore, the paths that minimize the “energy” in (3) will maximize the probability in (1). Now imagine that we can choose the potential function, $V(\mathbf{x})$, in such a way as to generate an invariant probability measure on \mathbf{R}^n . In other words, imagine that we can find a steady-state solution to equation (2). We can then project our stochastic process onto a nonlinear subspace of \mathbf{R}^n — i.e., onto an embedded Riemannian manifold — and examine the probability density induced on that subspace. Feynman’s heuristic picture of the relationship between (1) and (3) suggests that the subspaces of maximal probability will also be the subspaces of minimal energy, and the hope is that this will lead us to a solution to the clustering and coding problems in \mathbf{R}^n .

However, there are several problems with this model:

- First, it is well known that Feynman’s heuristic interpretation of formula (1) is mathematical nonsense, since there is no analogue of Lebesgue measure in an infinite-dimensional space. The relationship between (1) and (2) holds rigorously, as stated, if $\mathcal{W}_{\mathbf{x}}$ is Wiener measure, or *Brownian motion*, but there is still a gulf between (1) and (3). To interpret the integral in (3) as an expression in differential geometry, the paths $X_s \equiv X(s) \equiv X(s, \omega)$ must be continuous and differentiable. But, under Wiener measure, *with probability one*, the paths $X(t, \omega)$ are continuous but *nowhere differentiable*. Thus there is a fundamental clash between the geometric model and the probabilistic model. Stroock calls this “a fact which ... haunts every attempt to deal with Brownian paths,” [Str96], p. 140.
- Second, assuming that we can overcome our first problem, it is not a simple matter to project a stochastic process from \mathbf{R}^n onto an embedded Riemannian manifold. The mathematical problem itself has only been solved, in general, during the course

of the past 20 or 30 years, and it is now part of a subject known as *stochastic differential geometry*. See [EM89] or [Hsu02]. But the calculations are not trivial.

- Finally, it would be a mistake to assume that the Feynman-Kac formula can be used directly to generate a stochastic process, with a proper probabilistic interpretation. Instead, we will need a new potential function, $U(\mathbf{x})$, and we will need a further derivation from equations (1) and (2), in order to construct a stochastic process with an invariant probability measure. This also means that we will not be able to define our dissimilarity metric, directly, by minimizing the energy functional in formula (3).

In the remainder of this section, we will address these three problems, in reverse order. Our analysis will eventually lead us to a modification of the naive Feynman-Kac model, and to the definition of a dissimilarity metric which will achieve the goals articulated in Section 1.

2.1. A Stochastic Process with an Invariant Measure. To see the problem with the basic Feynman-Kac formula, it is helpful to rewrite (1) using an operator:

$$(4) \quad [\mathbf{P}_t f](\mathbf{x}) = \int_{\Omega} f(X_t) \exp \left[- \int_0^t V(X_s) ds \right] \mathcal{W}_{\mathbf{x}}(dX)$$

It turns out that $\mathbf{P}_t \mathbf{1} \neq \mathbf{1}$, which means that we cannot use this operator to construct a Markov process with a proper probabilistic interpretation. Another manifestation of the same problem is the fact that $V(\mathbf{x})$ has a natural interpretation as the “killing rate” for the process, i.e., the probability per unit of time that a path starting at $X_0 = \mathbf{x}$ will “die” by time δt . Thus the process “dissipates” as time goes by.

To fix this problem, we need a new potential function. If μ is a probability density that satisfies $\frac{1}{2}\Delta\mu - V(\mathbf{x})\mu = 0$, then

$$V(\mathbf{x}) = \frac{1}{2} \left(\frac{\Delta\mu}{\mu} \right) = \frac{1}{2} (\Delta \log \mu + |\nabla \log \mu|^2)$$

The first equality is trivial, and the second equality follows from a straightforward computation, e.g., by expanding $\Delta \log \mu$ in Cartesian coordinates. This equation suggests that we should work with a potential function $U(\mathbf{x})$ and define $V(\mathbf{x})$ as follows:

$$(5) \quad V(\mathbf{x}) = \frac{1}{2} (\Delta U(\mathbf{x}) + |\nabla U(\mathbf{x})|^2)$$

Now consider the following initial value problem:

$$(6) \quad \frac{\partial w}{\partial t} = \frac{1}{2} \Delta w + \nabla U(\mathbf{x}) \cdot \nabla w \quad \text{with } w(0, \cdot) = f$$

Lemma 1. *$w(t, \mathbf{x})$ is a solution to (6) if and only if $e^{U(\mathbf{x})}w(t, \mathbf{x})$ is a solution to (2) with initial value $u(0, \cdot) = e^U f$.*

Proof. By a straightforward computation, using the definition in (5) of $V(\mathbf{x})$ in terms of $U(\mathbf{x})$. \square

We now use both U and V to define a new operator:

$$(7) \quad [\mathbf{Q}_t f](\mathbf{x}) = \exp[-U(X_0)] \int_{\Omega} f(X_t) \exp \left[U(X_t) - \int_0^t V(X_s) ds \right] \mathcal{W}_{\mathbf{x}}(dX)$$

Theorem 1. *If U is bounded above and V is bounded below, and if $w(t, \mathbf{x})$ is a solution to (6) and $e^{U(\mathbf{x})}w(t, \mathbf{x})$ is also bounded, then $w(t, \mathbf{x})$ is equal to $[\mathbf{Q}_t f](\mathbf{x})$ as defined in (7). Furthermore, $\mathbf{Q}_t \mathbf{1} = \mathbf{1}$ for all $t \geq 0$, and $(\mathbf{Q}_t)_{t \geq 0}$ is a semigroup of operators which defines a Markov process on \mathbf{R}^n with the invariant measure $\mu = e^{U(\mathbf{x})}$.*

Proof. See Theorem 4.3.36 in [Str93] or Theorem 10.3.33 in [Str11]. \square

In the literature, (6) is known as a *diffusion equation* with a *drift* term ∇U . It is a nice feature of our formalism that this drift term is the gradient of a potential $U(\mathbf{x})$, and that the invariant measure turns out to be the exponential of the potential $U(\mathbf{x})$. For a numerical example, if $U(\mathbf{x})$ is a negative quadratic polynomial (which would be bounded above), then $V(\mathbf{x})$ would be a positive quadratic polynomial (which would be bounded below), and the invariant measure would be a Gaussian. See Section 5.1 below.

Sources: These results appear in [Str93], Section 4.3, but the analysis there uses a different definition of V in terms of U . In the second edition of his book, Stroock switches to the more natural definition in (5) above, but with the opposite sign. See [Str11], Section 10.3. Øksendal also uses this example, with the same definition of V and the same sign, in Exercises 8.15 and 8.16 of his text [Øks03].

2.2. Mapping a Diffusion to an Embedded Manifold. The equations in the previous section were all expressed in Cartesian coordinates, and the results would be different in a different coordinate system. For a simple example, if the standard 2-dimensional Laplacian were transformed into polar coordinates, it would acquire an additional first-order “drift” term. This is a problem if we want to map a diffusion from \mathbf{R}^n onto a nonlinear Riemannian manifold.

One approach to this problem is to analyze the diffusion by means of a *stochastic differential equation*, in two versions, one due to Itô, and one due to Stratonovich. We will write a 1-dimensional Itô process as:

$$X(t) = X(0) + \int_0^t \sigma(s, \omega) d\mathcal{B}(s, \omega) + \int_0^t b(s, \omega) ds$$

where the first integral is an Itô integral defined with respect to the Brownian motion $\mathcal{B}(t, \omega)$, and the second integral is an ordinary Riemann or Lebesgue integral. In differential notation, this would be:

$$(8) \quad dX(t) = \sigma(t, \omega) d\mathcal{B}(t, \omega) + b(t, \omega) dt$$

Extending this notation to n dimensions, let $\mathcal{B}_1(t, \omega), \dots, \mathcal{B}_d(t, \omega)$ be d independent Brownian motion processes, assume that $\mathbf{b}: \mathbf{R}^n \rightarrow \mathbf{R}^n$ and $\sigma: \mathbf{R}^n \rightarrow \mathbf{R}^{n \times d}$ are Lipschitz continuous, and define the n -dimensional Itô process as follows:

$$(9) \quad dX(t) = \begin{pmatrix} \sigma_1^1 & \dots & \sigma_d^1 \\ \vdots & & \vdots \\ \sigma_1^n & \dots & \sigma_d^n \end{pmatrix} \begin{pmatrix} d\mathcal{B}_1(t) \\ \vdots \\ d\mathcal{B}_d(t) \end{pmatrix} + \begin{pmatrix} b^1 \\ \vdots \\ b^n \end{pmatrix} dt$$

We want to construct a differential operator associated with this process. Setting $\mathbf{a} = \sigma\sigma^T$, define \mathcal{L} for all $f \in C^2(\mathbf{R}^n; \mathbf{R})$ by:

$$(10) \quad [\mathcal{L}f](\mathbf{x}) = \frac{1}{2} \sum_{i,j} a^{ij}(\mathbf{x}) \frac{\partial^2 f}{\partial x^i \partial x^j} + \sum_i b^i(\mathbf{x}) \frac{\partial f}{\partial x^i}$$

Theorem 2. *The operator \mathcal{L} defined in (10) is the infinitesimal generator of the n -dimensional Itô process given by (9).*

Proof. See Definition 7.3.1 and Theorem 7.3.3 in [Øks03]. □

Intuitively, σ is the “square root” of \mathbf{a} . Note also that, if $\mathbf{a} = \sigma\sigma^T$ is the identity matrix and $\mathbf{b} = \nabla U$, then (9) and (10) give us the same stochastic process in \mathbf{R}^n as does (6).

For our purposes, however, the Itô process has a defect: It is not invariant under coordinate transformations. This can be seen by an examination of *Itô’s formula*, which functions as a “chain rule” for the stochastic calculus, but with a second-order correction term. Let $F: \mathbf{R}^n \rightarrow \mathbf{R}$ be a function with continuous second-order partial derivatives. Then Itô’s formula asserts that:

$$dF(X(t)) = \sum_i \frac{\partial F(X(t))}{\partial x^i} dX_i(t) + \frac{1}{2} \sum_{i,j} \frac{\partial^2 F(X(t))}{\partial x^i \partial x^j} dX_i(t) dX_j(t)$$

See [Øks03], Chapter 4. An alternative is to use the Stratonovich integral, which cancels out the correction term. A common notational

device is to insert the symbol “ \circ ” in (9) to indicate that the stochastic integral is intended to be interpreted in the Stratonovich sense rather than the Itô sense. Using this notation, the equation for $dF(X(t))$ would be written as:

$$(11) \quad dF(X(t)) = \sum_i \frac{\partial F(X(t))}{\partial x^i} \circ dX_i(t)$$

in accordance with the usual rules of the Newton-Leibniz calculus. Since F could be an arbitrary coordinate transformation, the use of the Stratonovich formula in (11), instead of Itô’s formula, makes it possible to combine the stochastic calculus with the traditional constructs of Riemannian geometry.

Fortunately, the Itô integral and the Stratonovich integral can be developed in parallel, and it is possible to choose whichever version works best in a particular application. In the 1-dimensional case, we will write the Stratonovich version of a stochastic process as follows:

$$X(t) = X(0) + \int_0^t \sigma(s, \omega) \circ d\mathcal{B}(s, \omega) + \int_0^t \tilde{b}(s, \omega) ds$$

Notice the notation “ $\circ d\mathcal{B}(s, \omega)$ ” here, and the use of the function $\tilde{b}(s, \omega)$ instead of $b(s, \omega)$. Written as a differential, this would be:

$$(12) \quad dX(t) = \sigma(t, \omega) \circ d\mathcal{B}(t, \omega) + \tilde{b}(t, \omega) dt$$

Extending this notation to n dimensions, we have:

$$(13) \quad dX(t) = \begin{pmatrix} \sigma_1^1 & \dots & \sigma_d^1 \\ \vdots & & \vdots \\ \sigma_1^n & \dots & \sigma_d^n \end{pmatrix} \circ \begin{pmatrix} d\mathcal{B}_1(t) \\ \vdots \\ d\mathcal{B}_d(t) \end{pmatrix} + \begin{pmatrix} \tilde{b}^1 \\ \vdots \\ \tilde{b}^n \end{pmatrix} dt$$

Lemma 2. *The stochastic process defined by the Itô integral in (9) is identical to the process defined by the Stratonovich integral in (13) if and only if*

$$(14) \quad \tilde{b}^i = b^i - \frac{1}{2} \sum_{k=1}^d \sum_{j=1}^n \frac{\partial \sigma_k^i}{\partial x^j} \sigma_k^j$$

Proof. See [Str66] or [Itô75]. □

We thus have a simple mapping between the two formalisms, with the advantage that the stochastic differential equation in Stratonovich form is invariant under coordinate transformations.

Lemma 2 has an interesting consequence if we start out with the stochastic process given by (6). Recall that $\mathbf{a} = \sigma\sigma^T$ is the identity and $\mathbf{b} = \nabla U$ in this case. Suppose we satisfy the condition $\mathbf{a} = \mathbf{I}$ by

setting $\sigma = \mathbf{I}$. Then the second term in (14) vanishes, and $\tilde{\mathbf{b}} = \mathbf{b}$. However, if we subsequently apply a nonlinear coordinate transformation to our process, or project it onto a nonlinear subspace, then the Ito and Stratonovich equations will diverge, and we will want to use the Stratonovich equation from then on.

Let us now reinterpret the preceding analysis as a general property of vector fields. Define the column vectors

$$\mathbf{A}_0 = \begin{pmatrix} \tilde{b}^1 \\ \vdots \\ \tilde{b}^n \end{pmatrix} \quad \text{and} \quad \mathbf{A}_k = \begin{pmatrix} \sigma_k^1 \\ \vdots \\ \sigma_k^n \end{pmatrix} \quad \text{for } k = 1, \dots, d$$

and rewrite (13) as:

$$(15) \quad dX(t) = (\mathbf{A}_1 | \dots | \mathbf{A}_d) \circ d\mathcal{B}(t) + \mathbf{A}_0 dt$$

We will think of a vector field as a differential operator, essentially the *directional derivative* with respect to a given vector \mathbf{V} . Let us write this in shorthand notation as $\mathbf{V}\partial$. It then makes sense to talk about the “square” of a vector field, which we can define as the composition of the differential operator with itself: $(\mathbf{V}\partial)^2 = \mathbf{V}\partial \circ \mathbf{V}\partial$. Expanding this formula in a coordinate system, we have:

$$(16) \quad \left(\sum_i V^i \frac{\partial}{\partial x^i} \right) \circ \left(\sum_j V^j \frac{\partial}{\partial x^j} \right) = \sum_{i,j} V^i V^j \frac{\partial^2}{\partial x^i \partial x^j} + \sum_{i,j} \frac{\partial V^i}{\partial x^j} V^j \frac{\partial}{\partial x^i}$$

Now apply this equation to each of the vector fields $\mathbf{A}_k \partial$.

Theorem 3. *If \mathcal{L} is the differential operator associated with the stochastic process defined in (15), then*

$$(17) \quad \mathcal{L} = \frac{1}{2} \sum_{k=1}^d (\mathbf{A}_k \partial)^2 + \mathbf{A}_0 \partial$$

Proof. By a straightforward computation, using (10), (14) and (16). \square

With Theorem 3 as a guide, we can bypass the Itô or Stratonovich stochastic differential equations entirely, and work directly with vector fields. This is our second (but closely related) approach to the problem of mapping diffusions to embedded manifolds. Let $\mathbf{V}_0 \partial$ and $\mathbf{V}_k \partial$,

for $k = 1, \dots, d$, be arbitrary vector fields, and define the differential operator

$$(18) \quad \mathcal{L} = \frac{1}{2} \sum_{k=1}^d (\mathbf{V}_k \partial)^2 + \mathbf{V}_0 \partial.$$

This is known as the *Hörmander form* for the operator \mathcal{L} , and it, too, can be shown to be invariant under coordinate transformations. See [Hör67]. Thus \mathcal{L} works just as well in an arbitrary manifold \mathcal{M} as it does in \mathbf{R}^n endowed with Cartesian coordinates. The only condition that we need to impose to guarantee that \mathcal{L} , as defined in (18), gives us a nondegenerate diffusion in \mathcal{M} is to require that the vector fields $\{\mathbf{V}_1(\mathbf{x})\partial, \dots, \mathbf{V}_d(\mathbf{x})\partial\}$ span the tangent space on \mathcal{M} at \mathbf{x} . For these reasons, Stroock relies on the Hörmander formalism extensively in his book on the analysis of Brownian paths on Riemannian manifolds [Str00].

Sources: For the basic results on stochastic differential equations, using Itô’s formalism, the reader should consult [Øks03], but Øksendal’s text provides only a cursory treatment of Stratonovich’s formalism. The original paper by Stratonovich [Str66] is still very readable, but his theory was only given a solid mathematical foundation some years later by Itô [Itô75]. Chapter 8 of [Str03] is an excellent contemporary account of Stratonovich’s theory, set in a broader context.

2.3. Integral Curves and Martingales on Manifolds. There remains the problem that “haunts every attempt to deal with Brownian paths,” [Str96], p. 140. How do you reconcile the “smooth” curves of differential geometry with the “rough” paths that provide the support for Wiener measure? One answer, suggested by Stroock, emerges from a study of the relationship between the integral curves of a vector field and the concept of a *martingale*.

Let’s examine this idea, first, in the ordinary Euclidean space \mathbf{R}^n . Roughly speaking, a (continuous parameter) martingale M_t is a stochastic process which is “conditionally constant” in the sense that

$$\mathbb{E}[M_t \mid \mathcal{F}_s] = M_s \text{ for all } 0 \leq s \leq t,$$

where the conditional expectation \mathbb{E} is taken with respect to an nondecreasing family of sub- σ -algebras $\{\mathcal{F}_s\}_{s \geq 0}$ with the property that each M_t is \mathcal{F}_t -measurable. Since we are only considering probability spaces $(\Omega, \mathcal{F}, \mathbb{P})$ in which Ω is the set of continuous paths in \mathbf{R}^n and for which the σ -algebras \mathcal{F} and $\{\mathcal{F}_s\}_{s \geq 0}$ are fixed, we will suppress these references in our notation, and refer simply to a “martingale with respect

to \mathbb{P} ,” or a \mathbb{P} -martingale. We are interested in the relationship between martingales and differential operators.

Definition 1. Let \mathcal{L} be a second-order differential operator, and let $\mathbb{P}_{\mathbf{x}}$ be a probability measure on the space $C([0, \infty); \mathbf{R}^n)$ of all continuous paths in \mathbf{R}^n such that $\mathbb{P}_{\mathbf{x}}(X_0 = \mathbf{x}) = 1$. We say that $\mathbb{P}_{\mathbf{x}}$ solves the martingale problem for \mathcal{L} starting at \mathbf{x} if

$$M_t \equiv f(X_t) - \int_0^t [\mathcal{L}f](X_s) ds$$

is a $\mathbb{P}_{\mathbf{x}}$ -martingale for every $f \in C^\infty(\mathbf{R}^n; \mathbf{R})$.

Not surprisingly:

Lemma 3. If $\mathcal{L} = \frac{1}{2}\Delta$, then the Wiener measure $\mathcal{W}_{\mathbf{x}}$ solves the martingale problem for \mathcal{L} starting at \mathbf{x} .

Proof. See Corollary 7.1.20 and Remark 7.1.23 in [Str93]. \square

Let us now consider the operator $\mathcal{L} = \mathbf{b} \cdot \nabla$ and the integral equation:

$$(19) \quad Y_t = \mathbf{x} + \int_0^t \mathbf{b}(Y_s) ds, \quad 0 \leq t,$$

where $Y_t \equiv Y(t)$ is a continuous path in \mathbf{R}^n . An equivalent differential equation is:

$$(20) \quad \begin{aligned} Y'(t) &= \mathbf{b}(Y(t)) \\ Y(0) &= \mathbf{x} \end{aligned}$$

By the existence and uniqueness theorem for ordinary differential equations, (19) and (20) have a unique solution, which would commonly be referred to as the *integral curve* of the vector field \mathbf{b} starting at \mathbf{x} . Intuitively, an integral curve is a curve whose tangent is identical to the given vector field at each point. Note, too, that an integral curve is a “smooth” curve if \mathbf{b} is a smooth vector field. We have the following result:

Lemma 4. Let $\mathcal{L} = \mathbf{b} \cdot \nabla$, and let $\mathbb{P}_{\mathbf{x}}$ be the unit point mass concentrated on the solution to (19) or (20). Then $\mathbb{P}_{\mathbf{x}}$ solves the martingale problem for \mathcal{L} starting at \mathbf{x} .

Proof. See Exercise 7.1.32 in [Str93]. \square

We now put these two examples together, and consider the differential operator:

$$(21) \quad \mathcal{L} = \frac{1}{2}\Delta + \mathbf{b} \cdot \nabla$$

along with the stochastic process determined by the integral equation:

$$(22) \quad Y_t = X_t + \int_0^t \mathbf{b}(Y_s) ds, \quad 0 \leq t$$

In this equation, we are assuming that X_t is our original stochastic process with the usual Wiener measure $\mathcal{W}_{\mathbf{x}}$, and Y_t is a derived process with a derived probability measure.

Theorem 4. *Let \mathcal{L} be the differential operator given by (21), and let $\mathbb{Q}_{\mathbf{x}}$ be the probability measure determined by (22) when X_t is a stochastic process whose probability law is given by Wiener measure. Then $\mathbb{Q}_{\mathbf{x}}$ solves the martingale problem for \mathcal{L} starting at \mathbf{x} .*

Proof. See Theorem 7.3.10 in [Str93]. \square

Intuitively, these results show that a stochastic process defined by (6), or (9), or (13), has a “pure” diffusion part and a “pure” drift part, and the drift part follows the integral curve of the drift vector.

The preceding analysis is not confined to Euclidean \mathbf{R}^n , since a similar construction works when \mathcal{L} is given in Hörmander form by (18), see [ST94], and this means that all results can be replicated in an arbitrary Riemannian manifold, see [ST96]. The theory is explicated further in [Str00], where it serves as the foundation for Stroock’s construction and analysis of Brownian motion on a Riemannian manifold. Specifically, Section 2.2.1 of [Str00] includes a generalization of Lemma 4 above, and Theorem 2.40 of [Str00] is a generalization of Theorem 4.

3. PROTOTYPE CODING.

In discussing the mathematical background of the paper in the previous section, we were actually developing, implicitly, the main elements of our geometric and probabilistic models. The potential function, $U(\mathbf{x})$, and its gradient, $\nabla U(\mathbf{x})$, were introduced in connection with equations (6) and (7) and Theorem 1. Equation (6) is a diffusion equation with a drift term, $\nabla U(\mathbf{x})$, and it has an invariant probability density equal to $e^{U(\mathbf{x})}$, modulo a normalization factor. The stochastic process described by equation (6) can also be written as an Itô process, using equations (9) and (10) and Theorem 2, or it can be written in Stratonovich form, using equation (13) and Lemma 2. An alternative view of equation (6) is given by Stroock’s result, Theorem 4, on the relationship between integral curves and martingales on manifolds.

Recall that the main goal of our theory is to construct a lower-dimensional subspace of the original Euclidean space, \mathbf{R}^n , which is “optimal” in some sense. To be specific, let’s say that the subspace

should be a k -dimensional Riemannian manifold, embedded in \mathbf{R}^n , with a local coordinate system centered at $(0, 0, \dots, 0)$. We will use a form of *prototype coding* for the coordinate system, measuring the *distance* from the *origin* (i.e., the “prototype”) in $k - 1$ specified *directions*. Extending this coordinate system to all of \mathbf{R}^n , we can assume that these $k - 1$ coordinate directions have been chosen from among $n - 1$ coordinate directions in the full space. We will now follow the strategy suggested at the beginning of Section 2 for the naive Feynman-Kac model. Choose $U(\mathbf{x})$ and $\nabla U(\mathbf{x})$ so that the invariant probability density for the stochastic process given by equation (6) matches the density of our sample input data in \mathbf{R}^n . We can then *project* this stochastic process onto the embedded k -dimensional manifold, and examine the probability density induced on that manifold. The hope is that this procedure will lead us to the “best” k -dimensional coordinate system for the purpose of encoding our initial data.

How to do this? Our first step was described briefly in the text following Theorem 2 above. We start with (6): a diffusion equation with a drift vector, $\nabla U(\mathbf{x})$. We then write the differential operator associated with (6) in the form given by (10):

$$\mathcal{L} = \frac{1}{2} \sum_{i,j} a^{ij}(\mathbf{x}) \frac{\partial^2}{\partial x^i \partial x^j} + \sum_i b^i(\mathbf{x}) \frac{\partial}{\partial x^i}$$

by setting $\mathbf{a}(\mathbf{x})$ equal to the identity matrix, and setting $\mathbf{b}(\mathbf{x}) = \nabla U(\mathbf{x})$. By Theorem 2, \mathcal{L} is the infinitesimal generator of the n -dimensional Itô process given by (9):

$$dX(t) = \begin{pmatrix} \sigma_k^i(\mathbf{x}) \\ \vdots \\ \sigma_n^i(\mathbf{x}) \end{pmatrix} \begin{pmatrix} d\mathcal{B}_1(t) \\ \vdots \\ d\mathcal{B}_n(t) \end{pmatrix} + \begin{pmatrix} b^1(\mathbf{x}) \\ \vdots \\ b^n(\mathbf{x}) \end{pmatrix} dt$$

The choice of $\sigma(\mathbf{x})$ is arbitrary, as long as $\mathbf{a}(\mathbf{x}) = \sigma(\mathbf{x})\sigma(\mathbf{x})^T$ is the identity matrix, which means that $\sigma(\mathbf{x})$ must be an orthogonal transformation. These equations are expressed in Cartesian coordinates.

To implement the idea of prototype coding, suppose we are given a radial coordinate, ρ , and the directional coordinates $\theta^1, \theta^2, \dots, \theta^{n-1}$. For convenience, we will use the symbol Θ to refer to the entire sequence

of directional coordinates. Assume the existence of n coordinate transformation functions, with the usual properties:

$$\begin{aligned} x^1 &= \mathbf{x}^1(\rho, \theta^1, \theta^2, \dots, \theta^{n-1}) \\ x^2 &= \mathbf{x}^2(\rho, \theta^1, \theta^2, \dots, \theta^{n-1}) \\ &\dots \\ x^n &= \mathbf{x}^n(\rho, \theta^1, \theta^2, \dots, \theta^{n-1}) \end{aligned}$$

Let $\mathbf{J}(\rho, \Theta)$ denote the Jacobian matrix of these transformation functions. We want to represent our stochastic process in this new coordinate system, and to do so we need to convert the Itô equation, given by (9), into a Stratonovich equation in the form given by (13). We have two equalities:

$$(23) \quad dX(t) = \begin{pmatrix} \sigma_k^i(\mathbf{x}) \end{pmatrix} \circ \begin{pmatrix} d\mathcal{B}_1(t) \\ \vdots \\ d\mathcal{B}_n(t) \end{pmatrix} + \begin{pmatrix} \tilde{b}^1(\mathbf{x}) \\ \vdots \\ \tilde{b}^n(\mathbf{x}) \end{pmatrix} dt$$

$$(24) \quad dX(t) = \begin{pmatrix} \mathbf{J}(\rho, \Theta) \end{pmatrix} \circ \begin{pmatrix} dX_\rho(t) \\ dX_{\theta^1}(t) \\ \vdots \\ dX_{\theta^{n-1}}(t) \end{pmatrix}$$

The first equality is justified by Lemma 2. The second equality is justified by the Stratonovich formula for the “chain rule,” given by (11). The notation $dX_\rho(t), dX_{\theta^1}(t), \dots, dX_{\theta^{n-1}}(t)$, in the second equation, expresses the fact that $X_\rho(t), X_{\theta^1}(t), \dots$, and $X_{\theta^{n-1}}(t)$ are intended to represent the components of a new stochastic process defined on (ρ, Θ) .

We can now combine and solve equations (23) and (24) to obtain:

$$\begin{aligned} \begin{pmatrix} dX_\rho(t) \\ dX_{\theta^1}(t) \\ \vdots \\ dX_{\theta^{n-1}}(t) \end{pmatrix} &= \begin{pmatrix} \mathbf{J}(\rho, \Theta) \end{pmatrix}^{-1} \begin{pmatrix} \sigma_k^i(\mathbf{x}(\rho, \Theta)) \end{pmatrix} \circ \begin{pmatrix} d\mathcal{B}_1(t) \\ \vdots \\ d\mathcal{B}_n(t) \end{pmatrix} + \\ &\quad \begin{pmatrix} \mathbf{J}(\rho, \Theta) \end{pmatrix}^{-1} \begin{pmatrix} \tilde{b}^1(\mathbf{x}(\rho, \Theta)) \\ \vdots \\ \tilde{b}^n(\mathbf{x}(\rho, \Theta)) \end{pmatrix} dt \end{aligned}$$

We thus have a representation of our original stochastic process, in Stratonovich form, but expressed entirely in the new (ρ, Θ) coordinate

system. Note that the second term in this solution is just the transformation law for a contravariant vector, or a type $(1, 0)$ tensor.

Now consider the decomposition of a Stratonovich stochastic differential equation as in (15):

$$\begin{pmatrix} dX_\rho(t) \\ dX_{\theta^1}(t) \\ \vdots \\ dX_{\theta^{n-1}}(t) \end{pmatrix} = (\mathbf{A}_1 | \dots | \mathbf{A}_n) \circ d\mathcal{B}(t) + \mathbf{A}_0 dt$$

By matching the components of this equation with the components of the preceding equation, we can determine the vector fields $\mathbf{A}_0\partial$ and $\mathbf{A}_1\partial, \dots, \mathbf{A}_n\partial$. Then, applying Theorem 3 and expanding the expression inside (17), we can compute a new infinitesimal generator, \mathcal{L} , for our stochastic process, expressed again entirely in the (ρ, Θ) coordinate system. Finally, whatever our result might be, it can be written in the following form:

$$(25) \quad \mathcal{L} = \frac{1}{2} \sum_{i,j=0}^{n-1} \alpha^{ij}(\rho, \Theta) \frac{\partial^2}{\partial y^i \partial y^j} + \sum_{i=0}^{n-1} \beta^i(\rho, \Theta) \frac{\partial}{\partial y^i}$$

where $y^0 = \rho$ and $y^i = \theta^i$, for $i = 1, \dots, n-1$. (To distinguish this equation for \mathcal{L} from the \mathcal{L} we started out with, we have written the coefficients of the differential operators as $\alpha^{ij}(\rho, \Theta)$ and $\beta^i(\rho, \Theta)$ instead of $a^{ij}(\mathbf{x})$ and $b^i(\mathbf{x})$.) Note that this is the infinitesimal generator of an Itô process, but we derived it by an excursion through Stratonovich!

Before proceeding further, we need to analyze the (ρ, Θ) coordinate system. How is it defined? What are its properties? First, we want the radial coordinate, ρ , to follow the drift vector, $\nabla U(\mathbf{x})$. We have already seen how to do this. Suppose $\hat{\rho}(t)$ is the solution to the following differential equation, based on (20):

$$\begin{aligned} \hat{\rho}'(t) &= \nabla U(\hat{\rho}(t)) \\ \hat{\rho}(0) &= \mathbf{x}_0 \end{aligned}$$

In other words, $\hat{\rho}(t)$ is the *integral curve* of the vector field $\nabla U(\mathbf{x})$ starting at \mathbf{x}_0 . This is almost the construction that we want for our radial coordinate, but not quite. We will actually work with a generalization of the concept of an integral curve, known as an *integral manifold*. A one-dimensional integral manifold is, roughly speaking, just the image of an integral curve without the parametrization, and it always exists, for any vector field. Since we want to be able to alter the parametrization of $\hat{\rho}(t)$, arbitrarily, in order to choose a suitable

coordinate, ρ , the one-dimensional integral manifold is the device that we need.

For the directional coordinates, $\theta^1, \theta^2, \dots, \theta^{n-1}$, the obvious generalization would be an integral manifold of dimension $n - 1$, orthogonal to the integral manifold for ρ . But, for $k \geq 2$, a k -dimensional integral manifold exists if and only if certain conditions are satisfied, known as the *Frobenius integrability conditions*. Fortunately, as we will see, if we are looking for an integral manifold orthogonal to a vector field that is proportional to the gradient of a potential function, such as $\nabla U(\mathbf{x})$, then the Theorem of Frobenius gives us the results that we want. Our analysis here is based on the standard literature in differential geometry. See, e.g., [Spi99], Chapter 6; [BG68], Chapter 3; [AM77], Chapter 8. We will discuss these results in Section 4.

To summarize: At this point, we have a one-dimensional integral manifold for the ρ coordinate, and an orthogonal $n - 1$ dimensional integral manifold for the Θ coordinates. But we want to construct a *lower-dimensional* subspace by projecting our stochastic process onto a $k - 1$ dimensional subset of the coordinates $\theta^1, \theta^2, \dots, \theta^{n-1}$. Taken together with the ρ coordinate, we want this operation to give us an “optimal” k dimensional subspace. The mathematical device that we need is a Riemannian metric, $g_{ij}(\mathbf{x})$, which we will use to measure *dissimilarity* on the integral manifolds. And crucially: *the dissimilarity metric should depend on the probability measure*. Roughly speaking, the dissimilarity should be small in a region in which the probability density is high, and large in a region in which the probability density is low. We can then take the following steps:

- To find a principal axis for the ρ coordinate, we minimize the Riemannian distance, $g_{ij}(\mathbf{x})$, along the drift vector.
- To choose the principal directions for the $\theta^1, \theta^2, \dots, \theta^{k-1}$ coordinates, we diagonalize the Riemannian matrix, $(g_{ij}(\mathbf{x}))$, and select the eigenvectors associated with the $k - 1$ smallest eigenvalues.
- To compute the coordinate curves, we follow the geodesics of the Riemannian metric, $g_{ij}(\mathbf{x})$, in each of the $k - 1$ principal directions.

Thus, overall, we are minimizing dissimilarity, and maximizing probability. We will show how to do this, using concrete examples, in Sections 5.1 and 5.2 of this paper.

In the following section, we will see how to construct an integral manifold orthogonal to ∇U , and how to define a dissimilarity metric, $g_{ij}(\mathbf{x})$, with the desired properties. Because of the prominent role

played by the Riemannian dissimilarity metric in our theory, it is natural to describe it as a theory of *differential similarity*.

4. INTEGRAL MANIFOLDS IN \mathbf{R}^3 .

From this point on, for purposes of exposition, we will restrict our investigations from \mathbf{R}^n to \mathbf{R}^3 . We will see later (in Section 7) that this is not a limitation on the scope of the theory, since our results can easily be generalized again to \mathbf{R}^n . Instead, the restriction to three dimensions simplifies our calculations, and makes them easier to visualize, as we will see in Section 5.

Since we are now working in three-dimensional Euclidean space, we are primarily interested in two-dimensional integral manifolds. Is there a two-dimensional integral manifold orthogonal to the drift vector, ∇U ? Consider, first, a more general case. Suppose $\mathbf{G} = (P(\mathbf{x}), Q(\mathbf{x}), R(\mathbf{x}))$ represents the coordinates of a vector field that is defined but not equal to $(0, 0, 0)$ in some open region $\mathcal{D} \subseteq \mathbf{R}^3$.

Theorem 5. *There exists a two-dimensional integral manifold in \mathcal{D} with tangent plane everywhere orthogonal to \mathbf{G} if and only if*

$$\mathbf{G} \cdot (\nabla \times \mathbf{G}) = 0$$

Proof. See [BC64], Problem 29, p. 23; [Car71], pp. 97–98; [LR75], pp. 155–156. \square

Intuitively, this theorem states that \mathbf{G} must be orthogonal to its own “curl,” a condition that is satisfied if \mathbf{G} is proportional to the gradient of a scalar potential. Thus, any \mathbf{G} in the form $N(\mathbf{x})\nabla U(\mathbf{x})$ would work.

We still need a method to compute this integral manifold, however, and to define a curvilinear coordinate system on it. One approach is to choose basis vectors for a two-dimensional subspace of the tangent space at \mathbf{x} in the following form:

$$(26) \quad \begin{aligned} \mathbf{V}\partial &= f(\mathbf{x})\frac{\partial}{\partial x} + \frac{\partial}{\partial y} \\ \mathbf{W}\partial &= g(\mathbf{x})\frac{\partial}{\partial x} + \frac{\partial}{\partial z} \end{aligned}$$

Now compute: $\mathbf{V} \times \mathbf{W} = (f, 1, 0) \times (g, 0, 1) = (1, -f, -g)$. If $\mathbf{V} \times \mathbf{W}$ is proportional to \mathbf{G} , then \mathbf{G} is orthogonal to the plane containing both \mathbf{V} and \mathbf{W} , and conversely. So we can set:

$$\begin{aligned} \mathbf{V} \times \mathbf{W} &= \frac{1}{P(\mathbf{x})}\mathbf{G} = \frac{1}{P(\mathbf{x})}(P(\mathbf{x}), Q(\mathbf{x}), R(\mathbf{x})) \\ &= (1, -f(\mathbf{x}), -g(\mathbf{x})) \end{aligned}$$

and obtain the results $f(\mathbf{x}) = -Q(\mathbf{x})/P(\mathbf{x})$ and $g(\mathbf{x}) = -R(\mathbf{x})/P(\mathbf{x})$. If $G = \nabla U(\mathbf{x})$, then Theorem 5 applies. In this case, the vector fields given by $P(\mathbf{x})\mathbf{V} = (-Q(\mathbf{x}), P(\mathbf{x}), 0)$ and $P(\mathbf{x})\mathbf{W} = (-R(\mathbf{x}), 0, P(\mathbf{x}))$ provide what we want, namely, a basis for the tangent plane to the two-dimensional integral manifold that is everywhere orthogonal to the drift vector, ∇U .

This construction can also be justified directly by the Theorem of Frobenius. Geometrically, we interpret $\mathbf{V}\partial$ and $\mathbf{W}\partial$ as the basis vectors for a *tangent subbundle*, E , in \mathbf{R}^3 . (Historically, a tangent subbundle was called a “distribution,” but this term does not have the right connotations today.) We compute the *Lie bracket* of $\mathbf{V}\partial$ and $\mathbf{W}\partial$ as follows:

$$(27) \quad [\mathbf{V}\partial, \mathbf{W}\partial] = \mathbf{V}\partial \circ \mathbf{W}\partial - \mathbf{W}\partial \circ \mathbf{V}\partial \\ = \left[\frac{\partial g}{\partial y} - \frac{\partial f}{\partial z} + f(x, y, z) \frac{\partial g}{\partial x} - g(x, y, z) \frac{\partial f}{\partial x} \right] \frac{\partial}{\partial x}$$

Now the geometric version of the Theorem of Frobenius asserts that, if $[\mathbf{V}\partial, \mathbf{W}\partial]$ “belongs to” E whenever $\mathbf{V}\partial$ and $\mathbf{W}\partial$ “belong to” E , then E can be extended to a full integral manifold in \mathbf{R}^3 . But if $\mathbf{V}\partial$ and $\mathbf{W}\partial$ are defined by the equations in (26) and also “belong to” E , then $[\mathbf{V}\partial, \mathbf{W}\partial]$ “belongs to” E if and only if the bracketed expression on the right-hand side of (27) is identically zero. This leads to the following classical statement of the Frobenius integrability conditions as a system of partial differential equations:

$$\frac{\partial g}{\partial y} + f(x, y, z) \frac{\partial g}{\partial x} = \frac{\partial f}{\partial z} + g(x, y, z) \frac{\partial f}{\partial x}$$

As a further check on Theorem 5, we can verify by a direct computation that the preceding equation holds for $f(\mathbf{x})$ and $g(\mathbf{x})$, as defined previously, when $G = \nabla U(\mathbf{x})$.

To simplify the notation, let us absorb the factor $P(\mathbf{x})$ into the definition of the two tangential vector fields, and write:

$$\begin{aligned} \nabla U(\mathbf{x}) &= (P(\mathbf{x}), Q(\mathbf{x}), R(\mathbf{x})) \\ \mathbf{V}(\mathbf{x}) &= (-Q(\mathbf{x}), P(\mathbf{x}), 0) \\ \mathbf{W}(\mathbf{x}) &= (-R(\mathbf{x}), 0, P(\mathbf{x})) \end{aligned}$$

In this form, it is easy to see that $\nabla U(\mathbf{x})$ is orthogonal to both $\mathbf{V}(\mathbf{x})$ and $\mathbf{W}(\mathbf{x})$. Note also that $\mathbf{V}(\mathbf{x})$ and $\mathbf{W}(\mathbf{x})$ are not orthogonal to each other, although the vector fields $\mathbf{V}\partial = \mathbf{V}(\mathbf{x})/P(\mathbf{x})$ and $\mathbf{W}\partial = \mathbf{W}(\mathbf{x})/P(\mathbf{x})$ commute, as we have seen, when viewed as differential operators. Now one way to use these tangential vector fields is to compute a *global* (ρ, θ, ϕ) coordinate system. For example, we can compute the integral

curves of the vector field $\mathbf{V}(\mathbf{x})$ and use these for a coordinate called θ , and we can compute the integral curves of the vector field $\mathbf{W}(\mathbf{x})$ and use these for a coordinate called ϕ . Note that the θ coordinate curves will all lie in the global xy plane, and the ϕ coordinate curves will all lie in the global xz plane, if we take this approach.

But another approach is to use these vector fields to construct a *local* coordinate system. Any linear combination of $\mathbf{V}(\mathbf{x})$ and $\mathbf{W}(\mathbf{x})$ could be taken as one of the basis vectors for the tangent subbundle, and we can vary this linear combination as we move around the integral manifold. To implement this idea, it is useful to define a *Riemannian metric* on the integral manifold. The most natural way to do this is to define a metric tensor on all of \mathbf{R}^3 , using the inner products of $\nabla U(\mathbf{x})$, $\mathbf{V}(\mathbf{x})$ and $\mathbf{W}(\mathbf{x})$, in that order. We thus define:

$$\begin{pmatrix} g_{ij}(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} P^2(\mathbf{x}) + Q^2(\mathbf{x}) + R^2(\mathbf{x}) & 0 & 0 \\ 0 & P^2(\mathbf{x}) + Q^2(\mathbf{x}) & Q(\mathbf{x})R(\mathbf{x}) \\ 0 & Q(\mathbf{x})R(\mathbf{x}) & P^2(\mathbf{x}) + R^2(\mathbf{x}) \end{pmatrix}$$

To remain consistent with the coordinate notation introduced in Section 3, we let i and j range over 0, 1, 2, and we stipulate that $y^0 = \rho$, $y^1 = \theta$, $y^2 = \phi$. Since $P(\mathbf{x})$, $Q(\mathbf{x})$, $R(\mathbf{x})$, are the components of the drift vector, $\nabla U(\mathbf{x})$, and since the diffusion equation in which $\nabla U(\mathbf{x})$ appears has an invariant probability density that is determined by the exponential of the potential function, $U(\mathbf{x})$, it should be clear that $g_{ij}(\mathbf{x})$ has at least some of the properties that we have been looking for. We thus adopt this formula, provisionally, as the definition of our *dissimilarity metric*.

The matrix $(g_{ij}(\mathbf{x}))$ is not diagonal, in general, but it can easily be diagonalized. The eigenvectors are:

$$\xi_1 = \begin{pmatrix} 0 \\ -R(\mathbf{x}) \\ Q(\mathbf{x}) \end{pmatrix}, \quad \xi_2 = \begin{pmatrix} 0 \\ Q(\mathbf{x}) \\ R(\mathbf{x}) \end{pmatrix}, \quad \xi_3 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix},$$

and the corresponding eigenvalues are: $\lambda_1(\mathbf{x}) = P^2(\mathbf{x})$ and $\lambda_2(\mathbf{x}) = \lambda_3(\mathbf{x}) = P^2(\mathbf{x}) + Q^2(\mathbf{x}) + R^2(\mathbf{x})$. This analysis leads to a spectral decomposition of $(g_{ij}(\mathbf{x}))$ as:

$$\lambda_1(\mathbf{x}) \kappa(\mathbf{x}) \begin{pmatrix} 0 & 0 & 0 \\ 0 & R^2(\mathbf{x}) & -Q(\mathbf{x})R(\mathbf{x}) \\ 0 & -Q(\mathbf{x})R(\mathbf{x}) & Q^2(\mathbf{x}) \end{pmatrix} +$$

$$\lambda_2(\mathbf{x}) \kappa(\mathbf{x}) \begin{pmatrix} 0 & 0 & 0 \\ 0 & Q^2(\mathbf{x}) & Q(\mathbf{x})R(\mathbf{x}) \\ 0 & Q(\mathbf{x})R(\mathbf{x}) & R^2(\mathbf{x}) \end{pmatrix} + \lambda_2(\mathbf{x}) \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

where $\kappa(\mathbf{x}) = 1/(Q^2(\mathbf{x}) + R^2(\mathbf{x}))$. Obviously, the third term in this expression corresponds to the ρ coordinate. Since $\lambda_1(\mathbf{x})$ is always the smallest eigenvalue, we will use the eigenvector ξ_1 to determine the initial direction of the θ coordinate on the surface of the integral manifold orthogonal to $\nabla U(\mathbf{x})$. It is obvious, too, that ξ_1 and ξ_2 are orthogonal to each other, even though $\mathbf{V}(\mathbf{x})$ and $\mathbf{W}(\mathbf{x})$ are not.

The main application of our dissimilarity metric, however, is to compute *geodesics* on the surface of the integral manifold orthogonal to $\nabla U(\mathbf{x})$. Recall that any linear combination of $\mathbf{V}(\mathbf{x})$ and $\mathbf{W}(\mathbf{x})$ yields a vector in the tangent subbundle, E , and thus we can construct vector fields in E in the form $v(t)\mathbf{V}(\mathbf{x}) + w(t)\mathbf{W}(\mathbf{x})$ for arbitrary functions $v(t)$ and $w(t)$. For a geodesic, we are looking for a curve $\gamma(t)$ with values in \mathbf{R}^3 which minimizes the “energy” functional:

$$(28) \quad \frac{1}{2} \int_0^T \begin{pmatrix} v(t) & w(t) \end{pmatrix} \begin{pmatrix} g_{11}(\gamma(t)) & g_{12}(\gamma(t)) \\ g_{21}(\gamma(t)) & g_{22}(\gamma(t)) \end{pmatrix} \begin{pmatrix} v(t) \\ w(t) \end{pmatrix} dt$$

subject to the constraint:

$$(29) \quad \gamma'(t) = v(t)\mathbf{V}(\gamma(t)) + w(t)\mathbf{W}(\gamma(t))$$

This variational problem leads to a system of Euler-Lagrange equations for the curves $\gamma(t) = (x(t), y(t), z(t))$ and $(v(t), w(t))$, plus three Lagrange multipliers. For initial conditions, we specify $(x(0), y(0), z(0))$ and we use the smallest eigenvalue of $g_{ij}(x(0), y(0), z(0))$ to determine the initial value $(v(0), w(0))$. This is a complicated system of equations, but it can be solved numerically in *Mathematica*.

The preceding analysis was based on a global coordinate system centered on the x axis, since our initial vector fields were determined by the equations $P(\mathbf{x})\mathbf{V} = (-Q(\mathbf{x}), P(\mathbf{x}), 0)$ and $P(\mathbf{x})\mathbf{W} = (-R(\mathbf{x}), 0, P(\mathbf{x}))$. But we could also work with a coordinate system centered on the y axis, using the equations $Q(\mathbf{x})\mathbf{V} = (0, -R(\mathbf{x}), Q(\mathbf{x}))$ and $Q(\mathbf{x})\mathbf{W} = (Q(\mathbf{x}), -P(\mathbf{x}), 0)$, or the z axis, using $R(\mathbf{x})\mathbf{V} = (R(\mathbf{x}), 0, -P(\mathbf{x}))$ and $R(\mathbf{x})\mathbf{W} = (0, R(\mathbf{x}), -Q(\mathbf{x}))$. In fact, it is useful to be able to switch from one such coordinate system to another, as we move around the

integral manifold. Since $g_{00}(\mathbf{x}) = P^2(\mathbf{x}) + Q^2(\mathbf{x}) + R^2(\mathbf{x}) = |\nabla U(\mathbf{x})|^2$, it is obvious that the ρ coordinate is independent of the global coordinate system used to define it. But the same is true of $g_{ij}(\mathbf{x})$ when $i \neq 0$ and $j \neq 0$. To see this, let y^1 and y^2 denote the Θ coordinates centered on the x axis, and let \bar{y}^1 and \bar{y}^2 denote the Θ coordinates centered on the y axis. The Jacobian matrix of the coordinate transformation from \bar{y}^k to y^i can be computed as follows:

$$\begin{aligned} & \begin{pmatrix} \partial y^i / \partial \bar{y}^k \end{pmatrix} \\ &= \begin{pmatrix} \partial x / \partial \rho & -Q(\mathbf{x}) & -R(\mathbf{x}) \\ \partial y / \partial \rho & P(\mathbf{x}) & 0 \\ \partial z / \partial \rho & 0 & P(\mathbf{x}) \end{pmatrix}^{-1} \begin{pmatrix} \partial x / \partial \rho & 0 & Q(\mathbf{x}) \\ \partial y / \partial \rho & -R(\mathbf{x}) & -P(\mathbf{x}) \\ \partial z / \partial \rho & Q(\mathbf{x}) & 0 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & -R(\mathbf{x})/P(\mathbf{x}) & -1 \\ 0 & Q(\mathbf{x})/P(\mathbf{x}) & 0 \end{pmatrix} \end{aligned}$$

Now let $g_{ij}(\mathbf{x})$ and $\bar{g}_{kl}(\mathbf{x})$ denote the dissimilarity metric based on the y^i and \bar{y}^k coordinates, respectively. Restricting our attention to the 2×2 matrix for the Θ coordinates, we compute:

$$\begin{aligned} & \begin{pmatrix} \bar{g}_{11}(\mathbf{x}) & \bar{g}_{12}(\mathbf{x}) \\ \bar{g}_{21}(\mathbf{x}) & \bar{g}_{22}(\mathbf{x}) \end{pmatrix} \\ &= \begin{pmatrix} Q^2(\mathbf{x}) + R^2(\mathbf{x}) & P(\mathbf{x})R(\mathbf{x}) \\ P(\mathbf{x})R(\mathbf{x}) & P^2(\mathbf{x}) + Q^2(\mathbf{x}) \end{pmatrix} \\ &= \begin{pmatrix} -R(\mathbf{x})/P(\mathbf{x}) & -1 \\ Q(\mathbf{x})/P(\mathbf{x}) & 0 \end{pmatrix}^T \begin{pmatrix} g_{11}(\mathbf{x}) & g_{12}(\mathbf{x}) \\ g_{21}(\mathbf{x}) & g_{22}(\mathbf{x}) \end{pmatrix} \begin{pmatrix} -R(\mathbf{x})/P(\mathbf{x}) & -1 \\ Q(\mathbf{x})/P(\mathbf{x}) & 0 \end{pmatrix} \end{aligned}$$

But this is just an instantiation of the transformation law for a type (0,2) tensor:

$$\bar{g}_{kl}(\mathbf{x}) = \sum_{i,j=1}^2 \frac{\partial y^i}{\partial \bar{y}^k} g_{ij}(\mathbf{x}) \frac{\partial y^j}{\partial \bar{y}^l}$$

The same calculations obviously lead to the same results for all pairwise transformations among the three global coordinate systems. Thus, on a two-dimensional integral manifold, for a fixed ρ , the dissimilarity metric, $g_{ij}(\mathbf{x})$, is independent of the global coordinate system used to define it.

5. EXPERIMENTS WITH *Mathematica*.

To sharpen our intuitions, and before developing the theory of differential similarity any further, let's look at some experiments in \mathbf{R}^3 using the computational and graphical facilities of *Mathematica*. Section 5.1 is a comprehensive study of the Gaussian case, which is the one example that can be solved analytically. Section 5.2 then considers what we will refer to as the “curvilinear Gaussian” case. Here, we apply a quadratic potential function to the output of a cubic polynomial coordinate transformation, producing an example that cannot be solved analytically, but which still retains some degree of tractability. Finally, in Section 5.3, we put two “curvilinear Gaussians” together in a mixture distribution.

The source code for these examples is available in three *Mathematica* notebooks:

```
Gaussian.nb
CurvilinearGaussian.nb
BimodalCurvilinearGaussian.nb
```

5.1. The Gaussian Case. Consider, first, the case of a quadratic potential, for which most results can be obtained analytically in closed form. Define $U(\mathbf{x})$ as follows:

$$U(x, y, z) = -\frac{1}{2}(ax^2 + by^2 + cz^2)$$

Then the gradient is: $\nabla U(x, y, z) = (-ax, -by, -cz)$, and the derived potential $V(\mathbf{x})$ is:

$$V(x, y, z) = \frac{1}{2}(a^2x^2 + b^2y^2 + c^2z^2) - \frac{1}{2}(a + b + c)$$

(We can ignore the constant term.) It is well known that the Feynman-Kac formula, given by either (4) or (7), has a closed-form solution whenever $U(\mathbf{x})$ and $V(\mathbf{x})$ are quadratic polynomials. Furthermore, the invariant probability measure, $e^{U(\mathbf{x})}$, given by Theorem 1, is obviously a Gaussian. Adding in the normalization factor, the invariant probability density function is:

$$(30) \quad \sqrt{abc} (2\pi)^{-\frac{3}{2}} \exp \left[-\frac{1}{2}(ax^2 + by^2 + cz^2) \right]$$

Note that the covariance matrix in (30) is already in diagonalized form.

For a numerical example, set $a = 1, b = 2, c = 4$. Figure 1 then shows the surface defined by the equation $U(x, y, z) = -2$. Figure 2 shows a **StreamPlot** of the gradient vector field generated by $\nabla U(x, y, z)$ at $z = 0$. This picture makes sense, intuitively. Notice that the drift vector is

“transporting probability mass towards the origin,” to counteract the dissipative effects of the diffusion term in the stochastic process. If the system is in perfect balance, of course, we have an invariant probability measure, which in this case is a Gaussian.

The Gaussian case is simple enough that we can solve the differential equations explicitly in *Mathematica*, using `DSolve`. First, the integral curve of the vector field $\nabla U(\mathbf{x}) = (P(\mathbf{x}), Q(\mathbf{x}), R(\mathbf{x}))$ starting at $\mathbf{x}_0 = (x_0, y_0, z_0)$ is given by:

$$\hat{\rho}(t) = \begin{pmatrix} x_0 e^{-at} \\ y_0 e^{-bt} \\ z_0 e^{-ct} \end{pmatrix}$$

For the tangential vector fields, we will start with a global coordinate system centered on the x axis, so that $\mathbf{V}(\mathbf{x}) = (-Q(\mathbf{x}), P(\mathbf{x}), 0) = (by, -ax, 0)$ and $\mathbf{W}(\mathbf{x}) = (-R(\mathbf{x}), 0, P(\mathbf{x})) = (cz, 0, -ax)$. Then the integral curve of the vector field $\mathbf{V}\partial$ starting at $\mathbf{x}_1 = (x_1, y_1, z_1)$ is given by:

$$\hat{\theta}(t) = \begin{pmatrix} x_1 \cos \sqrt{ab} t + y_1 \sqrt{b/a} \sin \sqrt{ab} t \\ y_1 \cos \sqrt{ab} t - x_1 \sqrt{a/b} \sin \sqrt{ab} t \\ z_1 \end{pmatrix}$$

and the integral curve of $\mathbf{W}\partial$ starting at $\mathbf{x}_2 = (x_2, y_2, z_2)$ is given by:

$$\hat{\phi}(t) = \begin{pmatrix} x_2 \cos \sqrt{ac} t + z_2 \sqrt{c/a} \sin \sqrt{ac} t \\ y_2 \\ z_2 \cos \sqrt{ac} t - x_2 \sqrt{a/c} \sin \sqrt{ac} t \end{pmatrix}$$

Figure 3 shows the global coordinate system on a two-dimensional integral manifold that would be generated by these curves. Note that the θ coordinate curves lie in the xy plane, and the ϕ coordinate curves lie in the xz plane, as expected.

Given the curves $\hat{\rho}(t)$, $\hat{\theta}(t)$ and $\hat{\phi}(t)$, what does it mean to say that a point in \mathbf{R}^3 has the coordinates (ρ, θ, ϕ) ? We adopt the following conventions: Starting with the x axis as the *principal axis*, choose a “maximal” point $\mathbf{x}_0 = (x_0, 0, 0)$ and follow the curve $\hat{\rho}(t)$ towards the origin. There are two natural measures of distance along this curve: the Euclidean arc length, which in this case is just the value of the x -coordinate, and the Riemannian arc length, which is determined by our dissimilarity metric, $g_{ij}(\mathbf{x})$. Our choice here is to use the Euclidean arc length to specify the ρ coordinate. (We will subsequently see another role for the Riemannian arc length.) In the Gaussian case, therefore, ρ has the value $x_0 e^{-at}$, which ranges over the interval $(0, x_0]$ as t ranges from ∞ to 0. But by choosing a value for ρ , we are also choosing the

integral manifold on which $\hat{\theta}(t)$ and $\hat{\phi}(t)$ are defined. Therefore, to interpret the coordinates θ and ϕ , starting at $(\rho, 0, 0)$, we traverse the distance θ along the $\hat{\theta}(t)$ curve from the slice $\theta = 0$, and we traverse the distance ϕ along the $\hat{\phi}(t)$ curve from the slice $\phi = 0$, until we arrive at the point (ρ, θ, ϕ) . Note, too, that we can traverse the $\hat{\theta}(t)$ and $\hat{\phi}(t)$ curves in either order, and still arrive at the same point, as long as we remain within a neighborhood of $(\rho, 0, 0)$ in which these curves intersect. The black dots in Figure 3 may be helpful in visualizing this procedure.

Once again, the Gaussian case is simple enough that we can analyze the coordinate transformation from (ρ, θ, ϕ) to (x, y, z) , and derive an explicit expression for its Jacobian matrix. First, let $\vec{\theta}_s(\mathbf{x}) = \hat{\theta}_s(s)$ denote the *flow* of the vector field $\mathbf{V}\partial$ starting at \mathbf{x} , and similarly let $\vec{\phi}_t(\mathbf{x}) = \hat{\phi}_t(t)$ denote the flow of the vector field $\mathbf{W}\partial$ starting at \mathbf{x} . Applying the composition, $\vec{\theta}_s \circ \vec{\phi}_t$, of the flows $\vec{\theta}_s$ and $\vec{\phi}_t$ to the point $\mathbf{x} = (\rho, 0, 0)$, we obtain the following equations, for arbitrary s and t :

$$(31) \quad \begin{aligned} x &= \vec{x}(\rho, s, t) = \rho \cos \sqrt{ab} s \cos \sqrt{ac} t \\ y &= \vec{y}(\rho, s, t) = -\rho \sqrt{a/b} \sin \sqrt{ab} s \cos \sqrt{ac} t \\ z &= \vec{z}(\rho, s, t) = -\rho \sqrt{a/c} \sin \sqrt{ac} t \end{aligned}$$

By a simple calculation:

$$\frac{\partial}{\partial s} \begin{pmatrix} \vec{x}(\rho, s, t) \\ \vec{y}(\rho, s, t) \\ \vec{z}(\rho, s, t) \end{pmatrix} = \begin{pmatrix} b \vec{y}(\rho, s, t) \\ -a \vec{x}(\rho, s, t) \\ 0 \end{pmatrix}$$

In other words, $\partial/\partial s = \mathbf{V}\partial$. By another simple calculation, setting $s = 0$ in (31), we have:

$$\frac{\partial}{\partial t} \begin{pmatrix} \vec{x}(\rho, 0, t) \\ \vec{y}(\rho, 0, t) \\ \vec{z}(\rho, 0, t) \end{pmatrix} = \begin{pmatrix} -\rho \sqrt{ac} \sin \sqrt{ac} t \\ 0 \\ -a \rho \cos \sqrt{ac} t \end{pmatrix} = \begin{pmatrix} c \vec{z}(\rho, 0, t) \\ 0 \\ -a \vec{x}(\rho, 0, t) \end{pmatrix}$$

In other words, $\partial/\partial t = \mathbf{W}\partial$ at $s = 0$. But we can now show that, if we move along the $\mathbf{V}\partial$ coordinate curves from the slice $s = 0$ at $(\rho, 0, t)$ to a new slice, where $s \neq 0$, the components of $\mathbf{W}\partial$ do not change. This result follows from the commutative property of the coordinate vector fields: $\mathbf{V}\partial \circ \mathbf{W}\partial = \mathbf{W}\partial \circ \mathbf{V}\partial$. See, e.g., [BG68], Theorem 3.7.1, and [Spi99], Lemma 5.13. We can obtain a similar result if we reverse the composition of the flows $\vec{\theta}_s$ and $\vec{\phi}_t$, and apply $\vec{\phi}_t \circ \vec{\theta}_s$ to the point $\mathbf{x} = (\rho, 0, 0)$. In this case, we can compute $\partial/\partial t = \mathbf{W}\partial$ for all s and t , and $\partial/\partial s = \mathbf{V}\partial$ for $t = 0$. Again, if we move along the $\mathbf{W}\partial$ coordinate

curves from the slice $t = 0$ at $(\rho, s, 0)$ to a new slice, where $t \neq 0$, the components of $\mathbf{V}\partial$ do not change.

Now consider the coordinate transformation itself. Start anywhere on the slice $s = 0$ and follow the flow $\vec{\theta}_s$ to $s = \theta$. Separately, start anywhere on the slice $t = 0$ and follow the flow $\vec{\phi}_t$ to $t = \phi$. If we are within a sufficiently small neighborhood of $(\rho, 0, 0)$, these curves will intersect at some point (x, y, z) on the integral manifold for ρ . Thus, applying the inverse function theorem, we can recover the coordinate values $(\rho, \theta, \phi) = (\rho(x, y, z), \theta(x, y, z), \phi(x, y, z))$. See, e.g., [BG68], Theorem 3.7.1, or [BC64], Theorem 1.5. Finally, even though we may not be able to write down an explicit formula for the coordinate transformation, in general, we can see from the equations above that the Jacobian matrix of $(x, y, z) = (x(\rho, \theta, \phi), y(\rho, \theta, \phi), z(\rho, \theta, \phi))$ can be written explicitly as:

$$(32) \quad \mathbf{J}(\rho, \theta, \phi) = \begin{pmatrix} x(\rho, \theta, \phi)/\rho & b y(\rho, \theta, \phi) & c z(\rho, \theta, \phi) \\ y(\rho, \theta, \phi)/\rho & -a x(\rho, \theta, \phi) & 0 \\ z(\rho, \theta, \phi)/\rho & 0 & -a x(\rho, \theta, \phi) \end{pmatrix}$$

This is all we need to carry out the calculations described in Section 3, including the calculation of the coefficients $\alpha^{ij}(\rho, \theta, \phi)$ and $\beta^i(\rho, \theta, \phi)$ in Equation (25). We will analyze these results further in Section 6.

We have referred to the x axis in Figure 3 as the “principal axis” because of its correspondence to the results of Principal Component Analysis (PCA) in traditional linear statistics. For the Gaussian probability density given by (30), with $a = 1, b = 2, c = 4$, the first component identified by PCA would be the x axis, and the second component would be the y axis. Thus the “principal surface” would be defined by the xy plane, which corresponds to the (ρ, θ) surface in our curvilinear coordinate system. Figure 4 depicts this surface, with the ρ and θ coordinates illustrated. The maximal point on the principal axis is $(10, 0, 0)$, and the θ coordinate curves have been evenly spaced along the ρ coordinate curve from $(10, 0, 0)$ to $(0, 0, 0)$. Similarly, the ρ coordinate curves have been evenly spaced along the maximal θ coordinate curve, which passes through the point $(10, 0, 0)$. Clearly, the (ρ, θ) surface in Figure 4 encodes the same information as the xy plane does in traditional linear statistics.

What if our potential function, $U(\mathbf{x})$, while still a quadratic, was not already in diagonalized form? Figure 5 shows an example in which the potential function depicted in Figure 1 has been rotated through the angle $\pi/3$ around the line from $(0, 0, 0)$ to $(1, 1, 1)$. Under this rotation, the maximal point on the principal axis, $(10, 0, 0)$, would be displaced to the position $(20/3, -10/3, 20/3)$. If our basis vectors, $\mathbf{V}(\mathbf{x})$

and $\mathbf{W}(\mathbf{x})$, were also rotated in the same way, we could still compute closed form solutions to the differential equations, using `DSolve`, and this procedure would still give us explicit expressions for the functions $\hat{\rho}(t)$, $\hat{\theta}(t)$ and $\hat{\phi}(t)$, although these expressions would be more complex than they were previously. Continuing in this way, as before, we would eventually produce the (ρ, θ) surface shown in Figure 6.

But this calculation would only be possible if we knew, *a priori*, what the rotation was. Suppose we had no information about the rotation. And, to make the task even more challenging, suppose we were not able to rely on symbolic methods in *Mathematica*, such as `DSolve`, but had to use numerical methods, such as `NDSolve`. Could we still compute the (ρ, θ) surface shown in Figure 6?

We need to address a preliminary issue: When we were working with `DSolve` in the simple Gaussian case, we were able to compute an explicit expression for $\hat{\rho}(t)$ and convert it into a formula for the ρ coordinate measured in Euclidean arc length. Basically, we were constructing a new parametrization of $\hat{\rho}(t)$. This is not easy to do in the general case, however, because it would require us to invert the general formula for arc length. Fortunately, there is a simpler approach, which works very well using `NDSolve`. In place of the differential equation derived from (20), we use the normalized version:

$$\begin{aligned}\gamma'(t) &= \frac{\nabla U(\gamma(t))}{\|\nabla U(\gamma(t))\|} \\ \gamma(0) &= \mathbf{x}_0\end{aligned}$$

Since our tangent vector now has length 1, the integral curve that solves this equation will be parametrized by Euclidean arc length, but otherwise it will be identical to $\hat{\rho}(t)$. The formula for Riemannian arc length, using our dissimilarity metric, $g_{ij}(\mathbf{x})$, is also very simple when $\gamma(t)$ is defined in this way:

$$\int_0^T \sqrt{\begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} g_{ij}(\gamma(t)) \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}} dt = \int_0^T \sqrt{g_{00}(\gamma(t))} dt$$

Note that a similar normalized differential equation could be used with the vector fields $\mathbf{V}(\mathbf{x})$ and $\mathbf{W}(\mathbf{x})$, if we wanted to compute integral curves $\hat{\theta}(t)$ and $\hat{\phi}(t)$ parametrized by Euclidean arc length, although the formula for the Riemannian arc length would be different.

We are now ready to compute the (ρ, θ) surface in Figure 6, without knowledge of the rotation, and without the use of analytical methods in

Mathematica. We will start off with the basis vectors $\mathbf{V}(\mathbf{x})$ and $\mathbf{W}(\mathbf{x})$ centered on the x axis. There are three steps:

- (1) Find a principal axis for the ρ coordinate.

The basic idea is to find a point (x_0, y_0, z_0) at a fixed Euclidean distance from the origin, and an integral curve $\gamma(t)$ which solves the normalized differential equation for $\nabla U(\mathbf{x})$ starting at $\mathbf{x} = (x_0, y_0, z_0)$, and for which the Riemannian distance, $g_{ij}(\mathbf{x})$, measured along $\gamma(t)$ for a fixed interval, t , is minimal. In short, we are looking for the *least* Riemannian distance for a fixed Euclidean distance.

We use `NDSolve` to compute $\gamma(t)$, and we use `NIntegrate` to compute the Riemannian distance along $\gamma(t)$. `FindMinimum` then searches for the minimal point (x_0, y_0, z_0) satisfying these constraints. In our rotated Gaussian example, we can start the search at $(10, 0, 0)$ with the constraint that (x_0, y_0, z_0) must lie on the sphere $x^2 + y^2 + z^2 = 100$, and `FindMinimum` will return the value $(x_0, y_0, z_0) = (6.66666, -3.33335, 6.66667)$. This is a reasonably good match with the analytical value, $(x_0, y_0, z_0) = (20/3, -10/3, 20/3)$.

- (2) Determine the initial directions of the θ coordinate.

The basic idea is to work with the eigenvector, $\xi_1(\mathbf{x})$, associated with the smallest eigenvalue, $\lambda_1(\mathbf{x})$, of the dissimilarity matrix, $(g_{ij}(\mathbf{x}))$, at the point $\mathbf{x} = (x_0, y_0, z_0)$. For expository purposes, let's initially use the analytical value $(x_0, y_0, z_0) = (20/3, -10/3, 20/3)$. Then the smallest eigenvalue is $400/9$, and its associated eigenvector is $(0, 2, 1)$. However, the inner product of the basis vectors, $\mathbf{V}(\mathbf{x})$ and $\mathbf{W}(\mathbf{x})$, at (x_0, y_0, z_0) is negative, which means that we need to flip the sign of, say, $\mathbf{W}(\mathbf{x})$, to determine the initial directions of the θ coordinate. We thus set $(v(0), w(0)) = (2, -1)$, for one direction, and $(v(0), w(0)) = (-2, 1)$, for the opposite direction.

There is only a minor difference if we use the numerical value $(x_0, y_0, z_0) = (6.66666, -3.33335, 6.66667)$. We then set $(v(0), w(0)) = (1.99996, -1)$, for one direction, and, for the opposite direction, $(v(0), w(0)) = (-1.99996, 1)$.

- (3) Compute the geodesic curves of the (ρ, θ) coordinate system.

In the final step, we compute the geodesic curves that solve the variational problem given by (28) and (29), with the initial value $(x(0), y(0), z(0)) = (6.66666, -3.33335, 6.66667)$ and

with $(v(0), w(0))$ equal to either $(1.99996, -1)$ or $(-1.99996, 1)$. *Mathematica* has a `VariationalMethods` package which computes the Euler-Lagrange equations symbolically from the specification of a variational problem. We use this package, and then solve the resulting equations numerically with `NDSolve`.

There is one complication: `NDSolve` encounters a singularity when the value of $P(\mathbf{x})$ gets very small and the values of $v(t)$ and $w(t)$ grow very large. We will examine a solution to this problem in Section 5.2, below.

The end result of these three steps is a radial coordinate curve, ρ , and an outer transverse coordinate curve, θ , which match perfectly the curves in Figure 6, within the expected tolerance of a numerical approximation and up to the location of the numerical singularities.

This match between the numerical results and the analytical results provides some evidence that our techniques are working correctly. We will now apply these techniques to an example for which analytical results are not available.

5.2. The Curvilinear Gaussian. For this example, we start with a cubic polynomial: $C(t) = t^3 - t^2 - t$. We then define a cubic polynomial coordinate transformation from (x, y, z) to (u, v, w) as follows:

$$\begin{aligned} u &= u(x, y, z) = C(1.4 y) + 2x(y^2 + z^2) \\ v &= v(x, y, z) = C(1.2 z) + 2y(z^2 + x^2) \\ w &= w(x, y, z) = C(1.0 x) + 2z(x^2 + y^2) \end{aligned}$$

Finally, we define $U(\mathbf{x})$ as a quadratic potential function in the variables u, v and w :

$$U(x, y, z) = -\frac{1}{2}(a u(x, y, z)^2 + b v(x, y, z)^2 + c w(x, y, z)^2) * 10^{-6}$$

Thus $U(\mathbf{x})$ is a sixth-degree polynomial in x, y and z , and the gradient, $\nabla U(\mathbf{x})$, is a fifth-degree polynomial. There are no known closed-form solutions to the Feynman-Kac formula, given by either (4) or (7), when $U(\mathbf{x})$ and $V(\mathbf{x})$ are higher-order polynomials. However, it is possible to discretize the Feynman-Kac “path integral,” and obtain approximate numerical solutions. See, for example, [Lya04].

For a numerical example, set $a = 1, b = 2, c = 4$. Figure 7 then shows the surface defined by the equation $U(x, y, z) = -10$. Figure 8 shows a `StreamPlot` of the gradient vector field generated by $\nabla U(x, y, z)$ at $z = -10$. Figure 9 shows a stack of such stream plots, at the values $z = 10, z = 0$ and $z = -10$. Notice how the drift vector twists and

turns to counteract the dissipative effects of the diffusion term, and maintain an invariant probability measure.

Figure 10 is analogous to Figure 3 in the Gaussian case, and depicts the integral manifold that passes through the point $(20, 0, -10)$. The coordinate curves in Figure 10 are generated by a global coordinate system centered on the x axis, with $P(\mathbf{x})\mathbf{V} = (-Q(\mathbf{x}), P(\mathbf{x}), 0)$ and $P(\mathbf{x})\mathbf{W} = (-R(\mathbf{x}), 0, P(\mathbf{x}))$. These curves are thus analogous to the global θ and ϕ coordinate curves shown in Figure 3.

Finally, Figure 11 shows the (ρ, θ) surface computed by our numerical techniques, and analogous to the (ρ, θ) surface in Figure 6. As before, we start off with the basis vectors $\mathbf{V}(\mathbf{x})$ and $\mathbf{W}(\mathbf{x})$ centered on the x axis, and we proceed through three steps:

- (1) Find a principal axis for the ρ coordinate.

Again, we use `NDSolve`, `NIntegrate` and `FindMinimum`. We start the search at $(20, 0, -10)$, and impose the constraint that (x_0, y_0, z_0) must lie on the sphere $x^2 + y^2 + z^2 = 500$. We obtain the result: $(x_0, y_0, z_0) = (20.4316, 1.27953, -8.99505)$.

- (2) Determine the initial directions of the θ coordinate.

Working with the eigenvector $\xi_1(\mathbf{x})$ at the point $\mathbf{x} = (20.4316, 1.27953, -8.99505)$, we set $(v(0), w(0)) = (7.03133, -1)$ and, alternatively, $(v(0), w(0)) = (-7.03133, 1)$. The first pair of values will give us the θ coordinate curve in the clockwise direction, as seen from the vantage point of Figure 11; the second pair of values will give us the θ coordinate curve in the counterclockwise direction.

- (3) Compute the geodesic curves of the (ρ, θ) coordinate system.

Using the initial values calculated in steps (1) and (2), we construct the Euler-Lagrange equations for the variational problem given by (28) and (29), and we solve them using `NDSolve`. We encounter singularities in both coordinate curves, however, and we cannot extend the geodesics into the negative x and positive z quadrant.

These three steps generate the principal ρ coordinate curve and the outer θ coordinate curves shown in Figure 11, at least in the positive x and the negative z quadrant. The remaining θ coordinate curves in this quadrant are evenly spaced along the ρ coordinate, and the remaining ρ coordinate curves are evenly spaced along the outer θ coordinate, and they have been generated by the same three steps.

To complete Figure 11, we need another approach. One solution is to draw a second set of coordinate curves starting from the *antipodal point* in the opposite quadrant. Again, there are three steps:

- (1) Find a principal axis for the ρ coordinate.

To find our initial starting point, (x_0, y_0, z_0) , we used `NDSolve`, `NIntegrate` and `FindMinimum` to search for the least Riemannian distance for a fixed Euclidean distance. In fact, the Riemannian distance from $(20.4316, 1.27953, -8.99505)$ to the origin is 6.30863, but this distance is the same from any point on the integral manifold. For example, the Riemannian distance to the origin along the ρ coordinate curves at 315° , 45° , and 90° , is also 6.30863, although the Euclidean distance is different in each case. Thus, to find the antipodal point in the negative x and positive z quadrant, we search along the integral manifold at a constant Riemannian distance to find the *greatest* Euclidean distance. `FindMaximum` gives us the result that we want: $(x_1, y_1, z_1) = (-19.2034, -1.25676, 9.25639)$.

- (2) Determine the initial directions of the θ coordinate.

Working with the eigenvector $\xi_1(\mathbf{x})$ at the point (x_1, y_1, z_1) , we set $(v(0), w(0)) = (7.36377, -1)$ for the θ coordinate curve in the clockwise direction, and $(v(0), w(0)) = (-7.36377, 1)$ for the θ coordinate curve in the counterclockwise direction.

- (3) Compute the geodesic curves of the (ρ, θ) coordinate system.

Using the initial values calculated in steps (1) and (2), we construct the Euler-Lagrange equations for the variational problem given by (28) and (29), and we solve them using `NDSolve`. Although we encounter singularities again in both cases, we can see from a close inspection of Figure 11 that there remains only a small gap to the location of the singularities along the θ coordinate curves that were previously constructed from (x_0, y_0, z_0) .

These three steps generate the principal ρ coordinate curve and the outer θ coordinate curves shown in Figure 11, in the negative x and positive z quadrant. The remaining θ coordinate curves in this quadrant are evenly spaced along the ρ coordinate, and the remaining ρ coordinate curves are evenly spaced along the outer θ coordinate. The reader will note that the coordinate curves in both quadrants in Figure 11 match up reasonably well.

5.3. The Bimodal Curvilinear Gaussian. Finally, we consider a bimodal case. Figure 12 shows two copies of the curvilinear Gaussian defined in Section 5.2. One copy has been translated from $(0, 0, 0)$ to $(20, 20, -10)$. The other copy has been translated from $(0, 0, 0)$ to $(-20, -20, 10)$ and rotated by $\pi/2$ around a line parallel to the y -axis. But the probability density is a *mixture*. If $U_1(\mathbf{x})$ is the potential function for the first copy and $U_2(\mathbf{x})$ is the potential function for the second copy, then the invariant probability density is given by:

$$e^{U(\mathbf{x})} \simeq e^{U_1(\mathbf{x})} + e^{U_2(\mathbf{x})},$$

modulo an appropriate normalization factor. Figure 12 is actually showing the surface defined by the equation:

$$e^{U_1(x,y,z)} + e^{U_2(x,y,z)} = 0.0001$$

The advantage of this representation lies in the fact that our calculations for each copy will be almost independent of each other. Observe that the effective potential function for the mixture will be:

$$U(\mathbf{x}) \simeq \log(e^{U_1(\mathbf{x})} + e^{U_2(\mathbf{x})})$$

Thus the gradient of $U(\mathbf{x})$ in a neighborhood of $(20, 20, -10)$ will be almost identical to the gradient of $U_1(\mathbf{x})$ computed by itself, and the gradient of $U(\mathbf{x})$ in a neighborhood of $(-20, -20, 10)$ will be almost identical to the gradient of $U_2(\mathbf{x})$ computed by itself.

The mixture distribution thus provides a useful representation of *clusters*. Analyzing the situation, intuitively, in terms of our dissimilarity metric, the two clusters in Figure 12 will be exponentially far apart.

6. DIFFUSION COEFFICIENTS AND DISSIMILARITY METRICS

Recall the main results from Section 3: We started with a diffusion process represented by an Ito stochastic differential equation, in Cartesian coordinates; we transformed this into a Stratonovich equation in the coordinates (ρ, Θ) ; and we then converted this back into an Ito process characterized by a differential operator with coefficients $\alpha^{ij}(\rho, \Theta)$ and $\beta^i(\rho, \Theta)$. The one necessary ingredient was the Jacobian matrix of the coordinate transformation.

As an illustration, let's try a brute force solution of these equations in the simple Gaussian case discussed in Section 5.1. The Jacobian is given by equation (32). For ease of reference, here is equation (23),

rewritten for the three-dimensional coordinate system (ρ, θ, ϕ) :

$$dX(t) = \begin{pmatrix} \sigma_k^i(\mathbf{x}(\rho, \theta, \phi)) \end{pmatrix} \circ \begin{pmatrix} d\mathcal{B}_1(t) \\ d\mathcal{B}_2(t) \\ d\mathcal{B}_3(t) \end{pmatrix} + \begin{pmatrix} \tilde{b}^1(\mathbf{x}(\rho, \theta, \phi)) \\ \tilde{b}^2(\mathbf{x}(\rho, \theta, \phi)) \\ \tilde{b}^3(\mathbf{x}(\rho, \theta, \phi)) \end{pmatrix} dt$$

For the moment, we will assume that $(\sigma_k^i(\mathbf{x}(\rho, \theta, \phi)))$ is an orthogonal transformation, but otherwise arbitrary. Our procedure is to combine and solve equations (23) and (24), and then expand the result using Theorem 3. When we do so, we discover that the “sum of squares” inside equation (17) yields an expression consisting of 2679 terms! However, by using the fact that $(\sigma_k^i(\mathbf{x}(\rho, \theta, \phi)))$ is an orthogonal transformation, we can eliminate all terms in which the factors σ_k^i appear without derivatives. Furthermore, all the terms that include derivatives of σ_k^i are cancelled out by similar terms in the expansion of $\mathbf{A}_0 \partial$ inside equation (17). The net result is equation (25), in the following form:

$$\mathcal{L} = \frac{1}{2} \sum_{i,j=0}^2 \alpha^{ij}(\rho, \theta, \phi) \frac{\partial^2}{\partial y^i \partial y^j} + \sum_{i=0}^2 \beta^i(\rho, \theta, \phi) \frac{\partial}{\partial y^i}$$

where $y^0 = \rho$, $y^1 = \theta$ and $y^2 = \phi$. Thus, the exact choice we make for the transformation $(\sigma_k^i(\mathbf{x}(\rho, \theta, \phi)))$ turns out to be irrelevant. However, the diffusion coefficients $\alpha^{ij}(\rho, \theta, \phi)$ and the drift coefficients $\beta^i(\rho, \theta, \phi)$ are still very complex, and they do not provide much insight into the structure of the solution, even in the simple Gaussian case.

For more insight, let’s separate the ρ coordinate from the Θ coordinates. The basic idea of the (ρ, Θ) coordinate system was to align the ρ coordinate with the drift vector, $\nabla U(\mathbf{x})$, so that the trajectory of our stochastic process in the direction of the Θ coordinates would be orthogonal to the drift. The definition of our dissimilarity metric, $(g_{ij}(\mathbf{x}))$, also exhibited a strong separation between the ρ coordinate and the Θ coordinates. So there is a natural question here: What is the relationship between the representation of our stochastic process in Θ coordinates and the Θ submatrix of $(g_{ij}(\mathbf{x}))$?

The answer is well known in the case of pure Brownian motion, without drift. The earliest example is in [Str71] and [Itô75]. Stroock discovered that if you project Brownian motion in \mathbf{R}^3 onto the surface of a sphere of radius r centered at $(0, 0, 0)$, the differential operator of the resulting stochastic process, in spherical coordinates, (r, ϑ, φ) , is:

$$\mathcal{L} = \frac{1}{2} \frac{1}{r^2} \left(\frac{\partial^2}{\partial \vartheta^2} + \frac{1}{\sin^2 \vartheta} \frac{\partial^2}{\partial \varphi^2} + \frac{1}{\tan \vartheta} \frac{\partial}{\partial \vartheta} \right)$$

which is the spherical Laplacian divided by 2. This result can be generalized to an arbitrary Riemannian manifold, \mathcal{M} , embedded in \mathbf{R}^n . For any $f \in C^\infty(\mathcal{M}; \mathbf{R})$, the *Laplace-Beltrami* operator, $\Delta_{\mathcal{M}}$, is defined by:

$$\Delta_{\mathcal{M}} f = \operatorname{div}_{\mathcal{M}}(\operatorname{grad}_{\mathcal{M}} f)$$

in which the *divergence*, $\operatorname{div}_{\mathcal{M}}$, and the *gradient*, $\operatorname{grad}_{\mathcal{M}}$, can both be defined on \mathcal{M} independently of a coordinate system.

Theorem 6. *Let \mathcal{M} be an embedded submanifold of \mathbf{R}^n , and let $\Delta_{\mathcal{M}}$ be the Laplace-Beltrami operator on \mathcal{M} . Then*

$$\mathcal{L} = \frac{1}{2} \Delta_{\mathcal{M}}$$

is the differential operator of a Brownian motion process in \mathbf{R}^n that has been projected orthogonally onto \mathcal{M} .

Proof. The proof starts by showing that \mathcal{L} can always be written in Hörmander form without the $\mathbf{V}_0 \partial$ term. In particular, we can write:

$$\mathcal{L} = \frac{1}{2} \Delta_{\mathcal{M}} = \frac{1}{2} \sum_{k=1}^n (\Pi_{\mathcal{M}}(\mathbf{e}_k \partial))^2$$

where $(\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n)$ is an orthonormal basis for \mathbf{R}^n and $\Pi_{\mathcal{M}}$ is the orthogonal projection operator from the tangent bundle in \mathbf{R}^n onto the tangent bundle in \mathcal{M} . See Section 4.2.1 of [Str00] or Theorem 3.1.4 in [Hsu02]. From this result, it follows that we can construct a diffusion process on \mathcal{M} whose increments are precisely the projections, under $\Pi_{\mathcal{M}}$, of the increments of a Brownian motion process in \mathbf{R}^n , and whose differential operator is \mathcal{L} . For the details, see Theorem 4.37 in [Str00]. \square

The diffusion process constructed in Theorem 6 is known as *Brownian motion on \mathcal{M}* .

Let us now analyze the stochastic process defined by equation (6), or (9), or (13), projected onto the ρ and Θ coordinates separately. To simplify the calculations, we will initially focus our attention on the simple Gaussian case, in which $\nabla U(x, y, z) = (-ax, -by, -cz)$, and we will start with a construction borrowed from [Str71] and [Itô75], but

adapted to match this example. Consider the following matrix:

$$\begin{pmatrix} \pi_j^i(x, y, z) \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \frac{1}{|\nabla U(x, y, z)|^2} \begin{pmatrix} -ax \\ -by \\ -cz \end{pmatrix} \begin{pmatrix} -ax & -by & -cz \end{pmatrix}$$

in which the product in the second line should be interpreted as the multiplication of a 3×1 matrix times a 1×3 matrix, yielding a 3×3 matrix. It is easy to check that $(\pi_j^i(\mathbf{x}))$ is idempotent:

$$\begin{pmatrix} \pi_k^i(x, y, z) \end{pmatrix} \begin{pmatrix} \pi_j^k(x, y, z) \end{pmatrix} = \begin{pmatrix} \pi_j^i(x, y, z) \end{pmatrix}$$

and that it maps the vector $\nabla U(x, y, z)$ onto the origin:

$$\begin{pmatrix} \pi_j^i(x, y, z) \end{pmatrix} \begin{pmatrix} -ax \\ -by \\ -cz \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

Thus $(\pi_j^i(\mathbf{x}))$ is a *projection* onto the plane tangent to the integral manifold at (x, y, z) .

We now apply this projection operator to the right-hand side of equation (23), as rewritten above. First, we set σ equal to the identity matrix, so that $\tilde{\mathbf{b}} = \mathbf{b} = \nabla U$. (See the discussion following Lemma 2 in Section 2.2.) Then the projection operator $(\pi_j^i(\mathbf{x}))$ annihilates the second term in (23), and we are left with:

$$(33) \quad dX(t) = \begin{pmatrix} \pi_j^i(\mathbf{x}(\rho, \theta, \phi)) \end{pmatrix} \circ \begin{pmatrix} d\mathcal{B}_1(t) \\ d\mathcal{B}_2(t) \\ d\mathcal{B}_3(t) \end{pmatrix}$$

We now combine equation (33) with equation (24), and solve this system of equations to obtain:

$$(34) \quad \begin{pmatrix} dX_\rho(t) \\ dX_\theta(t) \\ dX_\phi(t) \end{pmatrix} = \begin{pmatrix} \mathbf{J}(\rho, \theta, \phi) \end{pmatrix}^{-1} \begin{pmatrix} \pi_j^i(\mathbf{x}(\rho, \theta, \phi)) \end{pmatrix} \circ \begin{pmatrix} d\mathcal{B}_1(t) \\ d\mathcal{B}_2(t) \\ d\mathcal{B}_3(t) \end{pmatrix}$$

As a verification that our calculations are on the right track, we note that the multiplication of the two matrices on the right-hand side of (34) produces a matrix in which the first row is identically zero. This means that $dX_\rho(t) = 0$, which is exactly the result that we want.

We now continue the procedure outlined in Section 3, applying Theorem 3 to equation (34), and expanding the “sum of squares” inside (17). This allows us to compute the coefficients $\alpha^{ij}(\rho, \theta, \phi)$ and $\beta^i(\rho, \theta, \phi)$ in (25). It turns out that $\alpha^{ij}(\rho, \theta, \phi) = 0$ whenever $i = 0$ or $j = 0$, which is what we would expect. For the remaining diffusion coefficients, we compute:

$$\begin{aligned}\alpha^{11}(\rho, \theta, \phi) &= \frac{a^2 x(\rho, \theta, \phi)^2 + c^2 z(\rho, \theta, \phi)^2}{a^2 x(\rho, \theta, \phi)^2 |\nabla U|^2} \\ \alpha^{22}(\rho, \theta, \phi) &= \frac{a^2 x(\rho, \theta, \phi)^2 + b^2 y(\rho, \theta, \phi)^2}{a^2 x(\rho, \theta, \phi)^2 |\nabla U|^2} \\ \alpha^{12}(\rho, \theta, \phi) &= \alpha^{21}(\rho, \theta, \phi) = -\frac{b c y(\rho, \theta, \phi) z(\rho, \theta, \phi)}{a^2 x(\rho, \theta, \phi)^2 |\nabla U|^2}\end{aligned}$$

Alternatively, we can write the nonzero diffusion coefficients as a 2×2 matrix:

$$\begin{pmatrix} \alpha^{11}(\rho, \theta, \phi) & \alpha^{12}(\rho, \theta, \phi) \\ \alpha^{21}(\rho, \theta, \phi) & \alpha^{22}(\rho, \theta, \phi) \end{pmatrix} = \frac{1}{a^2 x(\rho, \theta, \phi)^2} \times \left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \frac{1}{|\nabla U|^2} \begin{pmatrix} -b y(\rho, \theta, \phi) \\ -c z(\rho, \theta, \phi) \end{pmatrix} \begin{pmatrix} -b y(\rho, \theta, \phi) & -c z(\rho, \theta, \phi) \end{pmatrix} \right)$$

It turns out also that the drift coefficient $\beta^0(\rho, \theta, \phi) = 0$, as we would expect, and for the other drift coefficients we compute:

$$\begin{aligned}\beta^1(\rho, \theta, \phi) &= \frac{b y(\rho, \theta, \phi)}{2 a x(\rho, \theta, \phi) |\nabla U|^2} \times \\ &\quad \left((b + c) + \frac{1}{|\nabla U|^2} \begin{pmatrix} a^2 x(\rho, \theta, \phi) \\ b^2 y(\rho, \theta, \phi) \\ c^2 z(\rho, \theta, \phi) \end{pmatrix} \cdot \nabla U \right) \\ \beta^2(\rho, \theta, \phi) &= \frac{c z(\rho, \theta, \phi)}{2 a x(\rho, \theta, \phi) |\nabla U|^2} \times \\ &\quad \left((b + c) + \frac{1}{|\nabla U|^2} \begin{pmatrix} a^2 x(\rho, \theta, \phi) \\ b^2 y(\rho, \theta, \phi) \\ c^2 z(\rho, \theta, \phi) \end{pmatrix} \cdot \nabla U \right)\end{aligned}$$

Keep in mind that these are the coefficients for the first-order terms $\partial/\partial\theta$ and $\partial/\partial\phi$.

For comparison, we will now compute the Laplace-Beltrami operator for the simple Gaussian case, using our Riemannian dissimilarity

metric, $g_{ij}(\rho, \theta, \phi) = g_{ij}(\mathbf{x}(\rho, \theta, \phi))$, on the two-dimensional integral manifold given by the Theorem of Frobenius. In a local coordinate system, the Laplace-Beltrami operator is usually written as follows:

$$\Delta_{\mathcal{M}} f = \frac{1}{\sqrt{G}} \sum_{j=1}^n \frac{\partial}{\partial y^j} \left(\sqrt{G} \sum_{i=1}^n g^{ij}(\mathbf{y}) \frac{\partial f}{\partial y^i} \right)$$

where G is the determinant of the matrix $(g_{ij}(\mathbf{y}))$ and $(g^{ij}(\mathbf{y}))$ is its inverse. Alternatively, we can expand the expression inside the parentheses, and write \mathcal{L} in the form of equation (25):

$$\begin{aligned} \mathcal{L} &= \frac{1}{2} \Delta_{\mathcal{M}} = \frac{1}{2} \sum_{i,j=1}^n g^{ij}(\mathbf{y}) \frac{\partial^2}{\partial y^i \partial y^j} + \sum_{i=1}^n h^i(\mathbf{y}) \frac{\partial}{\partial y^i}, \\ \text{with } h^i(\mathbf{y}) &= \frac{1}{2\sqrt{G}} \sum_{j=1}^n \frac{\partial (\sqrt{G} g^{ij}(\mathbf{y}))}{\partial y^j} \end{aligned}$$

When we do the calculations in the simple Gaussian case, with $n = 2$, we discover that the diffusion coefficients are identical:

$$(\alpha^{ij}(\rho, \theta, \phi)) = (g^{ij}(\rho, \theta, \phi))$$

and the drift coefficients are similar, but not identical:

$$\begin{aligned} h^1(\rho, \theta, \phi) &= \frac{b y(\rho, \theta, \phi)}{2 a x(\rho, \theta, \phi) |\nabla U|^2} \times \\ &\quad \left((a + b + c) + \frac{1}{|\nabla U|^2} \begin{pmatrix} a^2 x(\rho, \theta, \phi) \\ b^2 y(\rho, \theta, \phi) \\ c^2 z(\rho, \theta, \phi) \end{pmatrix} \cdot \nabla U \right) \\ h^2(\rho, \theta, \phi) &= \frac{c z(\rho, \theta, \phi)}{2 a x(\rho, \theta, \phi) |\nabla U|^2} \times \\ &\quad \left((a + b + c) + \frac{1}{|\nabla U|^2} \begin{pmatrix} a^2 x(\rho, \theta, \phi) \\ b^2 y(\rho, \theta, \phi) \\ c^2 z(\rho, \theta, \phi) \end{pmatrix} \cdot \nabla U \right) \end{aligned}$$

In fact, there is a simple relationship between the coefficients $\beta^i(\rho, \theta, \phi)$ and $h^i(\rho, \theta, \phi)$:

$$\begin{aligned} (35) \quad \beta^1(\rho, \theta, \phi) - h^1(\rho, \theta, \phi) &= - \frac{b y(\rho, \theta, \phi)}{2 x(\rho, \theta, \phi) |\nabla U|^2} \\ \beta^2(\rho, \theta, \phi) - h^2(\rho, \theta, \phi) &= - \frac{c z(\rho, \theta, \phi)}{2 x(\rho, \theta, \phi) |\nabla U|^2} \end{aligned}$$

Is there an explanation for these results?

The key is to recognize that the stochastic process defined by equation (6), or (9), or (13), is *not* Brownian motion. Brownian motion in \mathbf{R}^n dissipates, and does not generate an invariant probability measure. Thus the projection of Brownian motion onto a Riemannian manifold, \mathcal{M} , would dissipate as well. But the stochastic process defined by equation (6), when projected onto the manifold, \mathcal{M} , would not dissipate, in general. This difference must be reflected in the drift coefficients for $\partial/\partial\theta$ and $\partial/\partial\phi$, as shown by equation (35).

Figure 13 shows the “drift correction vector field” generated by equation (35) on one quadrant of the integral manifold through $(10, 0, 0)$, for the simple Gaussian case. The magnitude of the vector field is coded by color, with red indicating that the length of the vector is near zero. Keep in mind that we are looking at the *difference* between the two vector fields, $(\beta^i(\rho, \theta, \phi))$ and $(h^i(\rho, \theta, \phi))$, as defined by equation (35). The vector fields themselves are oriented (approximately) in the opposite direction, but they have different magnitudes.

We have presented detailed calculations for the simple Gaussian case, so that our results would be easy to visualize. But the same calculations work for the general case, $\nabla U(\mathbf{x}) = (P(\mathbf{x}), Q(\mathbf{x}), R(\mathbf{x}))$. The projection operator, $(\pi_j^i(\mathbf{x}(\rho, \theta, \phi)))$, and the Jacobian matrix, $(\mathbf{J}(\rho, \theta, \phi))$, can be defined in the same way, and the computational procedure from Section 3, applying Theorem 3 and expanding equation (17), still goes through. The expansion of the Laplace-Beltrami operator for the general dissimilarity metric, $g_{ij}(\rho, \theta, \phi) = g_{ij}(\mathbf{x}(\rho, \theta, \phi))$, also goes through. We end up, again, with diffusion coefficients that are identical:

$$\begin{aligned} (\alpha^{ij}(\rho, \theta, \phi)) &= (g^{ij}(\rho, \theta, \phi)) = \frac{1}{P^2(\mathbf{x}(\rho, \theta, \phi))} \times \\ &\left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \frac{1}{|\nabla U|^2} \begin{pmatrix} Q(\mathbf{x}(\rho, \theta, \phi)) \\ R(\mathbf{x}(\rho, \theta, \phi)) \end{pmatrix} \begin{pmatrix} Q(\mathbf{x}(\rho, \theta, \phi)) & R(\mathbf{x}(\rho, \theta, \phi)) \end{pmatrix} \right) \end{aligned}$$

and drift coefficients that differ by a single, but more complex, term:

$$\begin{aligned} \beta^1(\rho, \theta, \phi) - h^1(\rho, \theta, \phi) &= \\ &- \frac{1}{2 P^2 |\nabla U|^2} \left(P \left(P \frac{\partial Q}{\partial x} - Q \frac{\partial P}{\partial x} \right) + R \left(R \frac{\partial Q}{\partial x} - Q \frac{\partial R}{\partial x} \right) \right) \end{aligned}$$

$$\beta^2(\rho, \theta, \phi) - h^2(\rho, \theta, \phi) =$$

$$- \frac{1}{2 P^2 |\nabla U|^2} \left(P \left(P \frac{\partial R}{\partial x} - R \frac{\partial P}{\partial x} \right) + Q \left(Q \frac{\partial R}{\partial x} - R \frac{\partial Q}{\partial x} \right) \right)$$

Note that $\partial P / \partial x$ is the only partial derivative in these drift correction equations which is nonzero in the case $\nabla U(x, y, z) = (-ax, -by, -cz)$. Thus, for the simple Gaussian case, we can easily verify that the coefficients of the drift correction vector field reduce to the two terms:

$$- \frac{b y(\rho, \theta, \phi)}{2 x(\rho, \theta, \phi) |\nabla U|^2} \quad \text{and} \quad - \frac{c z(\rho, \theta, \phi)}{2 x(\rho, \theta, \phi) |\nabla U|^2}$$

in agreement with equation (35).

7. FUTURE WORK

The theory of differential similarity combines a stochastic model with a geometric model, and it works because there is a common mathematical object in both models: the gradient, $\nabla U(\mathbf{x})$, of a potential function, $U(\mathbf{x})$. In the stochastic model, $\nabla U(\mathbf{x})$ is the drift vector, which guarantees the existence of an invariant probability measure. In the geometric model, $\nabla U(\mathbf{x})$ guarantees the existence of an orthogonal integral manifold. We have seen, in Section 6, that there is a theoretical connection between these two models, in which $\nabla U(\mathbf{x})$ plays a crucial role, and we have seen the practical consequences of this connection in the computational examples in Sections 5.1 and 5.2.

The main deficiency in the theory, as presented in this paper, is the restriction of the geometric model to the three-dimensional case. We imposed this restriction to simplify the calculations, and to make it easy to visualize the examples in *Mathematica*. But the theory is not inherently limited to three dimensions. Theorem 5 in Section 4 was written using the vector cross product and the “curl,” which is a three-dimensional concept, but it is actually a special case of a general result in \mathbf{R}^n which follows from the dual version of the Theorem of Frobenius, expressed in terms of differential forms. It follows that $\mathbf{V}\partial$ and $\mathbf{W}\partial$, our basis vectors for the tangent subbundle in \mathbf{R}^3 , can be generalized to \mathbf{R}^n . In particular, if $\nabla U(\mathbf{x}) = (P_0(\mathbf{x}), P_1(\mathbf{x}), \dots, P_{n-1}(\mathbf{x}))$, we can

write:

$$\begin{aligned}
\nabla U(\mathbf{x}) &= (P_0(\mathbf{x}), P_1(\mathbf{x}), P_2(\mathbf{x}), \dots, P_{n-2}(\mathbf{x}), P_{n-1}(\mathbf{x})) \\
\mathbf{V}_1(\mathbf{x}) &= (-P_1(\mathbf{x}), P_0(\mathbf{x}), 0, \dots, 0, 0) \\
\mathbf{V}_2(\mathbf{x}) &= (-P_2(\mathbf{x}), 0, P_0(\mathbf{x}), \dots, 0, 0) \\
&\dots \\
\mathbf{V}_{n-2}(\mathbf{x}) &= (-P_{n-2}(\mathbf{x}), 0, 0, \dots, P_0(\mathbf{x}), 0) \\
\mathbf{V}_{n-1}(\mathbf{x}) &= (-P_{n-1}(\mathbf{x}), 0, 0, \dots, 0, P_0(\mathbf{x}))
\end{aligned}$$

It is straightforward to verify that $\nabla U(\mathbf{x})$ is orthogonal to each $\mathbf{V}_i(\mathbf{x})$, and that the tangent subbundle spanned by $\{\mathbf{V}_i \partial = \mathbf{V}_i(\mathbf{x})/P_0(\mathbf{x})\}$ satisfies the Frobenius integrability conditions. Thus the remaining results in Section 4 can be generalized as well. We will discuss these generalizations in a forthcoming paper, which has the working title: “Differential Similarity in Higher-Dimensional Spaces: Theory and Applications.”

Once the theory is extended to higher dimensions, it will become apparent that there are various connections to recent work in *manifold learning*, as described in Section 1. Work in this area tends to follow either a geometric approach or a probabilistic approach, but not both. Examples of the geometric approach include: [TSL00] [RS00] [BN03] [DG03]. Belkin and Niyogi [BN03], for example, work with the eigenvectors of the graph Laplacian and the eigenfunctions of the Laplace-Beltrami operator, and show that the solution to these eigenproblems yields an “optimal” embedding of a low-dimensional manifold into a higher-dimensional space, but their arguments are geometric rather than probabilistic. Examples of the probabilistic approach include: [HR02] [TB99] [CSP⁺10]. Tipping and Bishop [TB99] work with a mixture of low-dimensional Gaussians embedded in a higher-dimensional space, each with its own mean and covariance matrix, and they use the EM algorithm to estimate the parameters of this model. Chen, et al., [CSP⁺10] adopt a similar model, along with the assumption that the Gaussian mixture covers a low-dimensional manifold, and they estimate both the number of components in the mixture and the dimensionality of the subspaces, using Bayesian techniques. But neither paper makes use of the geometric structure of the embedded manifold.

One exception to this dichotomy between geometric and probabilistic approaches is a paper by Lee and Wasserman [LW10], which has some interesting connections to the present work. The paper starts out by defining a Markov chain on \mathbf{R}^n with a transition kernel $\Omega_\epsilon(\mathbf{x}, \cdot)$ which gives preference to nearby points, \mathbf{y} , that have a high probability density, $p(\mathbf{y})$. This kernel is then used to define the one-step diffusion

operator, A_ϵ , and its m -step version, $A_{\epsilon,m}$. The authors then construct a continuous time operator: $\mathbf{A}_t = \lim_{\epsilon \rightarrow 0} A_{\epsilon, t/\epsilon}$. The analogous mathematical object in our theory would be the operator \mathbf{Q}_t in equation (7). Lee and Wasserman are primarily interested in the eigenfunctions of $A_{\epsilon,m}$ and \mathbf{A}_t , which have applications to various *spectral clustering* problems, following the work of Belkin and Niyogi [BN03] and others. They also use $\Omega_\epsilon(\mathbf{x}, \cdot)$ to define a *diffusion distance*, $D_{\epsilon,m}^2(\mathbf{x}, \mathbf{z})$, and its continuous time version, $\mathbf{D}_t^2(\mathbf{x}, \mathbf{z})$, but there does not seem to be a straightforward relationship between this distance and our dissimilarity metric, $g_{ij}(\mathbf{x})$. The paper concludes with several examples that demonstrate the utility of these concepts.

The other important contribution of Lee and Wasserman [LW10] is their analysis of the statistical estimators for the population quantities, \mathbf{A}_t and \mathbf{D}_t^2 . This is essential future work for our theory as well, since we need to apply our model to real data in order to fully validate it.

One of the most important applications of manifold learning is to the field known as *deep learning* [BLPL06] [HOT06] [RPCL06]. In deep learning, a neural network with many layers (hence, “deep”) is trained in two stages: the first stage uses an unsupervised learning algorithm to construct a set of “features” bottom up; the second stage applies a supervised learning algorithm to the feature hierarchy, top down. In fact, the paper by Rifai, et al. [RDV⁺11], cited and quoted in Section 1, is an example of just such an approach. In their experiment, the authors extract a tangent plane at each training point using an unsupervised technique (CAE), and then train their network in a supervised manner using a (20 year old) technique that was designed to be sensitive to tangent directions (MTC). They write:

To the best of our knowledge this is the first time that the implicit relationship between an unsupervised learned mapping and the tangent space of a manifold is rendered explicit and successfully exploited for the training of a classifier.

To combine manifold learning and deep learning in this way, it is necessary to make a decision about the network interface: What kind of representation of the data should we use as the output of one layer and the input of the next? There are reasons to think that *prototype coding* (see Section 3) with our Riemannian dissimilarity metric, $g_{ij}(\mathbf{x})$, will turn out to be a good choice, but this is obviously an important question on our agenda for future work.

REFERENCES

- [AM77] Louis Auslander and Robert E. MacKenzie. *Introduction to Differentiable Manifolds*. Dover Publications, 1977.
- [BC64] Richard L. Bishop and Richard J. Crittenden. *Geometry of Manifolds*. Pure and applied mathematics. Academic Press, 1964.
- [BG68] Richard L. Bishop and Samuel I. Goldberg. *Tensor Analysis on Manifolds*. Macmillan, 1968.
- [BLPL06] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. In *Advances in Neural Information Processing Systems*, volume 19, pages 153–160, 2006.
- [BN03] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.
- [Car71] Henri Cartan. *Differential Forms*. Dover Books on Mathematics Series. Dover Publications, 1971.
- [CSP⁺10] M. Chen, J. Silva, J. W. Paisley, C. Wang, D. B. Dunson, and L. Carin. Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: Algorithm and performance bounds. *IEEE Transactions on Signal Processing*, 58(12):6140–6155, 2010.
- [DG03] D. Donoho and C. Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of National Academy of Sciences*, 100:5591–5596, 2003.
- [DH73] Richard O. Duda and Peter E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York, 1973.
- [DHS01] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*, chapter 10: Unsupervised Learning and Clustering. Wiley & Sons, Inc., New York, 2nd edition, 2001.
- [EM89] Michel Emery and Paul A. Meyer. *Stochastic Calculus in Manifolds*. World Publishing Company, 1989.
- [Fey48] R. P. Feynman. Space-time approach to non-relativistic quantum mechanics. *Rev. Mod. Phys.*, 20:367–387, 1948.
- [Hör67] L. Hörmander. Hypoelliptic second order differential equations. *Acta Mathematica*, 119:147–171, 1967.
- [HOT06] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006.
- [HR02] G. E. Hinton and S. T. Roweis. Stochastic neighbor embedding. In *Advances in Neural Information Processing Systems*, volume 15, pages 833–840, 2002.
- [Hsu02] Elton P. Hsu. *Stochastic Analysis on Manifolds*. Contemporary Mathematics. American Mathematical Society, 2002.
- [Itô75] K. Itô. Stochastic differentials. *Applied Mathematics & Optimization*, 1(4):374–381, 1975.
- [Kac49] M. Kac. On distributions of certain Wiener functionals. *Trans. Amer. Math. Soc.*, 65:1–13, 1949.
- [LR75] David Lovelock and Hanno Rund. *Tensors, Differential Forms, and Variational Principles*. Pure and Applied Mathematics. Wiley, 1975.
- [LW10] A. B. Lee and L. Wasserman. Spectral connectivity analysis. *Journal of the American Statistical Association*, 105(491):1241–1255, 2010.

- [Lya04] A. Lyasoff. Path integral methods for parabolic partial differential equations with examples from computational finance. *Mathematica Journal*, 9(2):399–422, 2004.
- [Øks03] Bernt K. Øksendal. *Stochastic Differential Equations: An Introduction With Applications*. Springer, sixth edition, 2003.
- [RDV⁺11] S. Rifai, Y. Dauphin, P. Vincent, Y. Bengio, and X. Muller. The manifold tangent classifier. In *Advances in Neural Information Processing Systems*, volume 24, pages 2294–2302, 2011.
- [RPCL06] M. Ranzato, C. Poultney, S. Chopra, and Y. LeCun. Efficient learning of sparse representations with an energy-based model. In *Advances in Neural Information Processing Systems*, volume 19, pages 1137–1144, 2006.
- [RS00] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.
- [Spi99] Michael Spivak. *A Comprehensive Introduction to Differential Geometry*, volume 1. Publish or Perish, third edition, 1999.
- [ST94] D. W. Stroock and S. Taniguchi. Diffusions as integral curves, or Stratonovich without Itô. In *The Dynkin Festschrift. Markov processes and their applications. In celebration of Eugene B. Dynkin's 70th birthday*, pages 333–369. Boston, MA: Birkhäuser, 1994.
- [ST96] D. W. Stroock and S. Taniguchi. Diffusions as integral curves on manifolds and Lie groups. In *Probability theory and mathematical statistics. Lectures presented at the semester held in St. Petersburg, Russia, March 2–April 23, 1993*, pages 219–226. Amsterdam: Gordon and Breach Publishers, 1996.
- [Str66] R. L. Stratonovich. A new representation for stochastic integrals and equations. *SIAM Journal on Control*, 4(2):362–371, 1966.
- [Str71] D. W. Stroock. On the growth of stochastic integrals. *Z. Wahrscheinlichkeitstheor. Verw. Geb.*, 18:340–344, 1971.
- [Str93] Daniel W. Stroock. *Probability Theory: An Analytic View*. Cambridge University Press, 1993.
- [Str96] D. W. Stroock. Gaussian measures in traditional and not so traditional settings. *Bulletin (New Series) of the American Mathematical Society*, 33(2):135–155, 1996.
- [Str00] Daniel W. Stroock. *An Introduction to the Analysis of Paths on a Riemannian Manifold*. Mathematical Surveys and Monographs. American Mathematical Society, 2000.
- [Str03] Daniel W. Stroock. *Markov Processes from K. Itô's Perspective*. Annals of Mathematics Studies. Princeton University Press, 2003.
- [Str11] Daniel W. Stroock. *Probability Theory: An Analytic View*. Cambridge University Press, second edition, 2011.
- [TB99] M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analyzers. *Neural Computation*, 11(2):443–482, 1999.
- [TSL00] J. B. Tenenbaum, V. Silva, and J. C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, 2000.

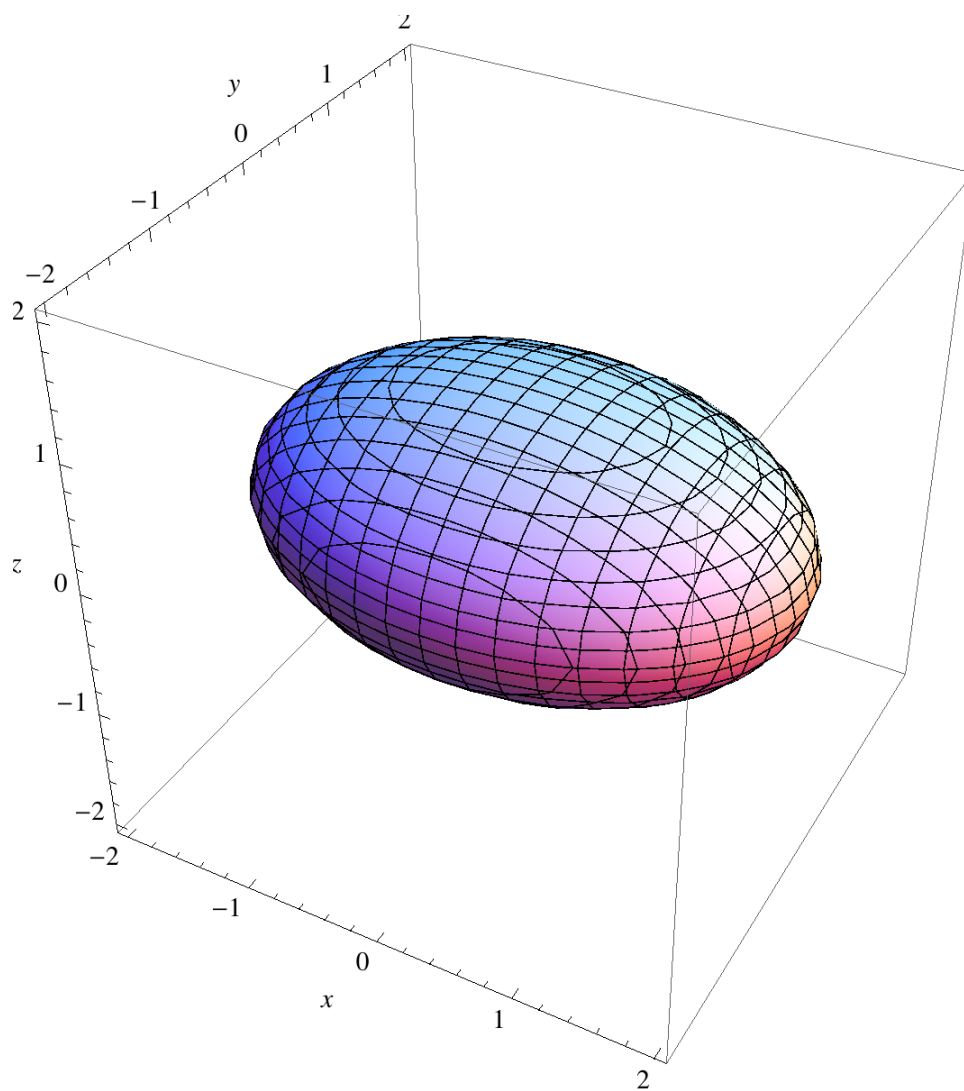


FIGURE 1. Contour Plot for the Surface of a Quadratic Potential.

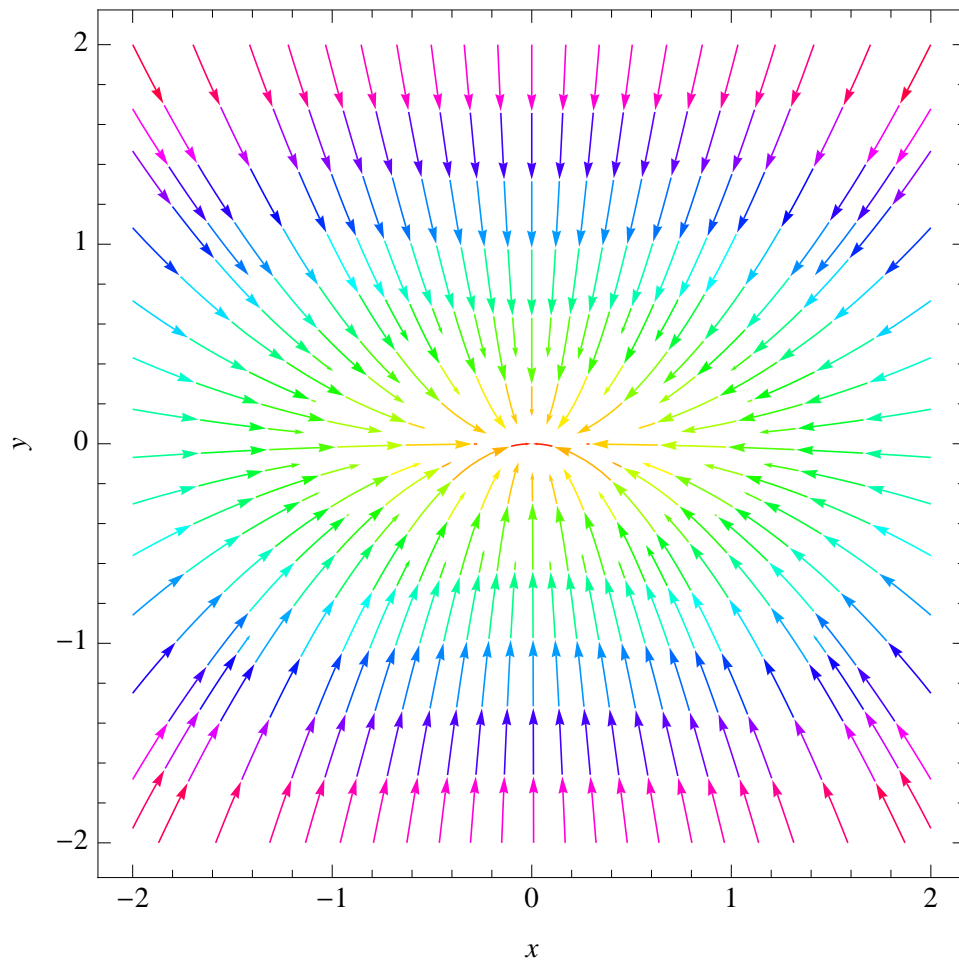


FIGURE 2. Gradient Vector Field for a Quadratic Potential at $z = 0$.

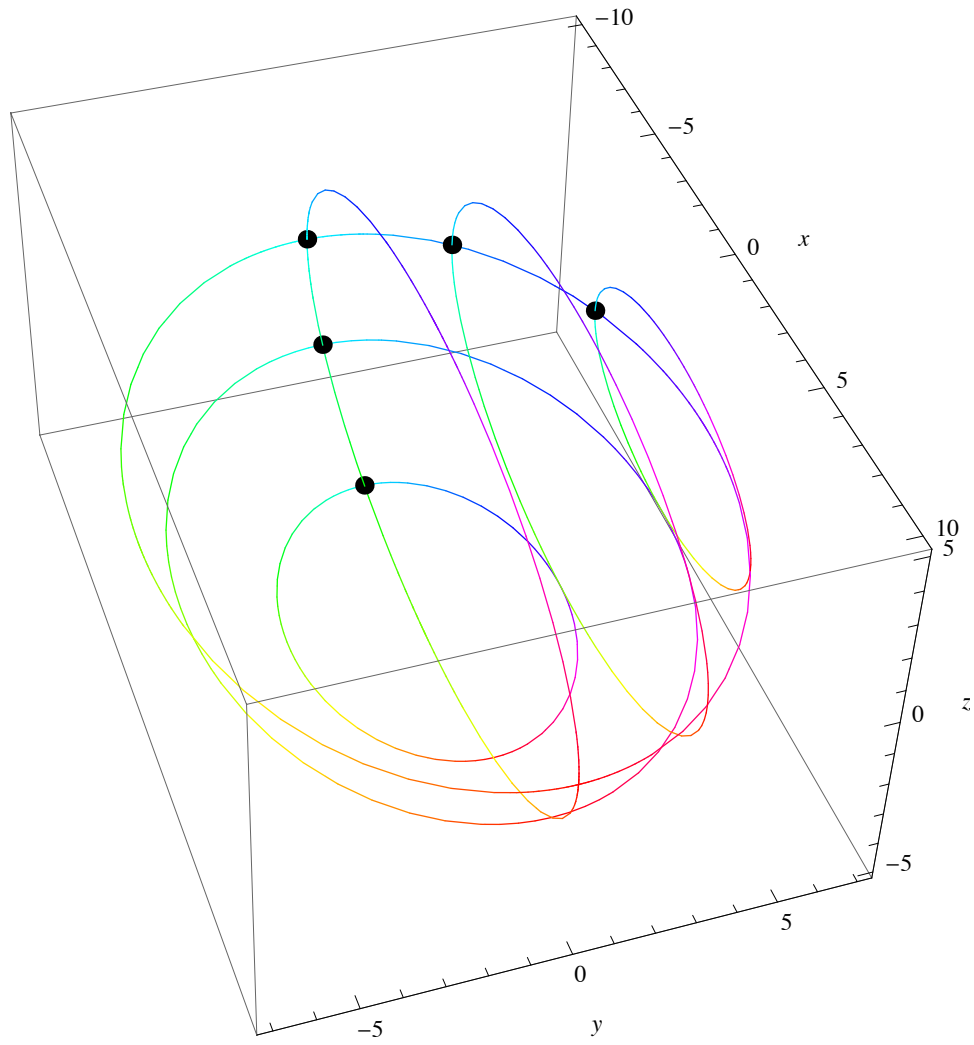


FIGURE 3. An Integral Manifold with a Curvilinear Co-ordinate System for a Quadratic Potential.

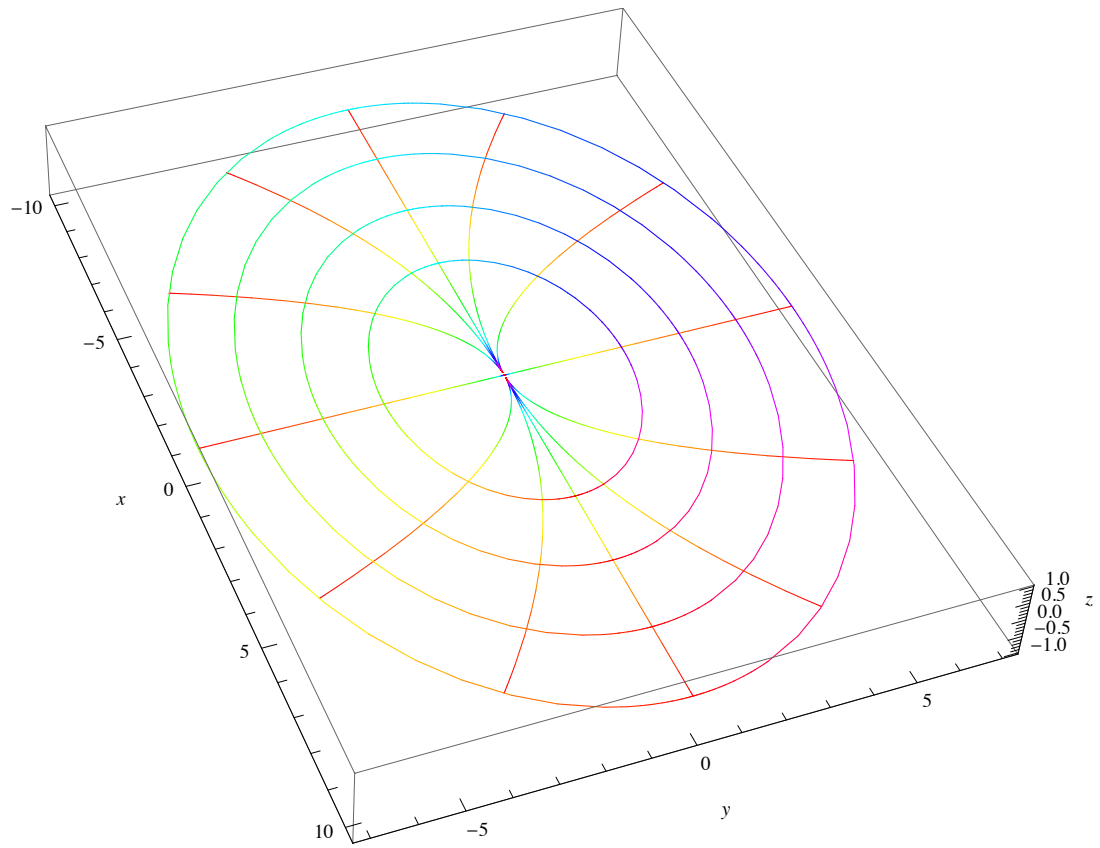


FIGURE 4. Curvilinear Coordinate System for the ρ, θ Surface of a Quadratic Potential.

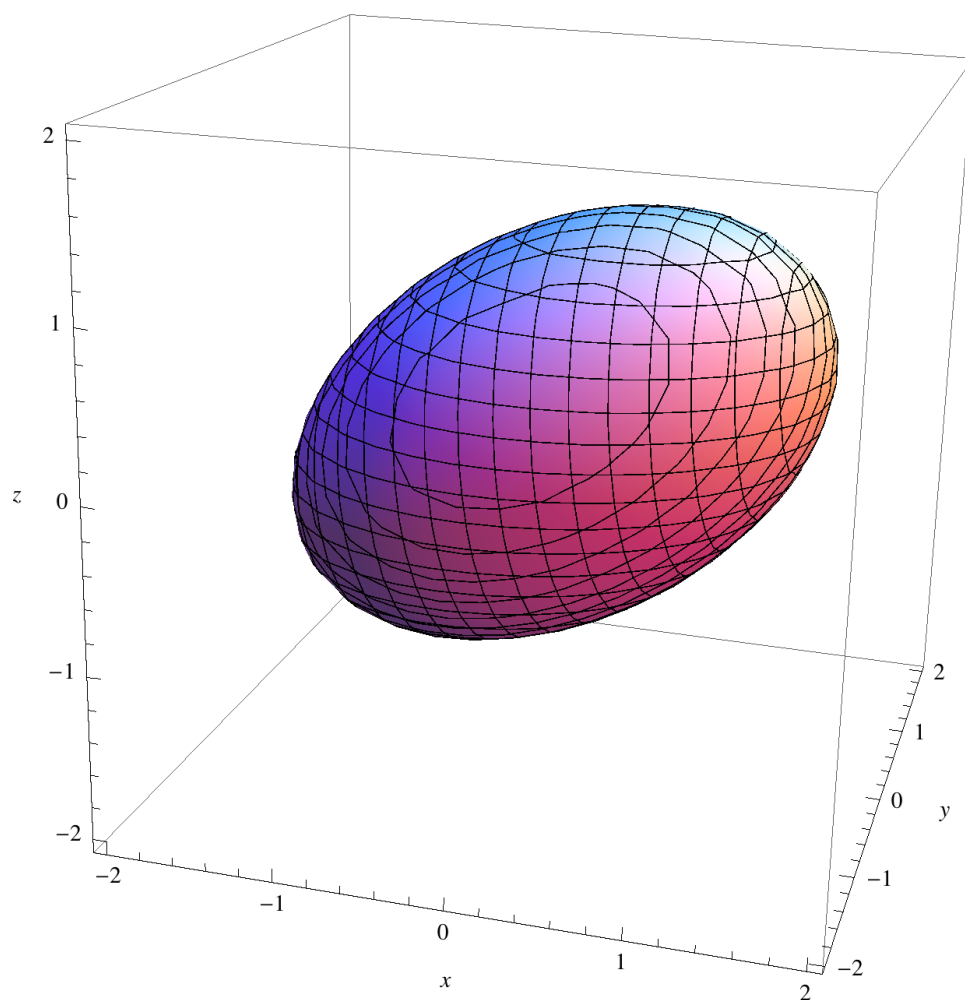


FIGURE 5. Contour Plot for the Surface of a Rotated Quadratic Potential.

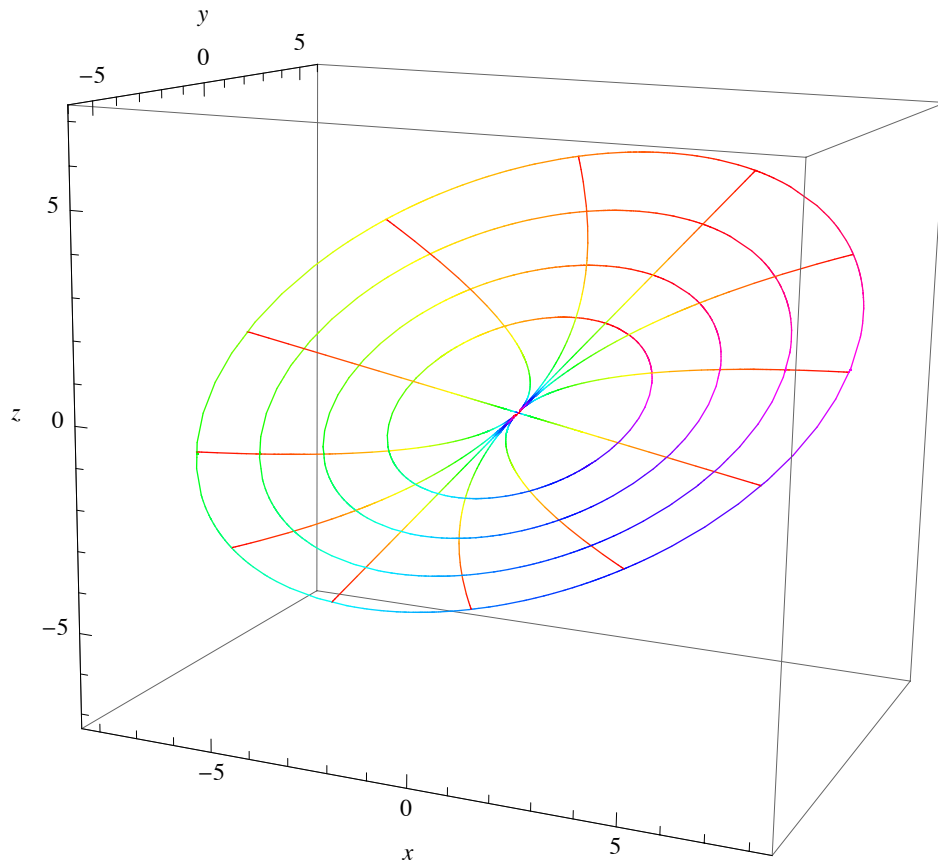


FIGURE 6. Curvilinear Coordinate System for the ρ, θ Surface of a Rotated Quadratic Potential.

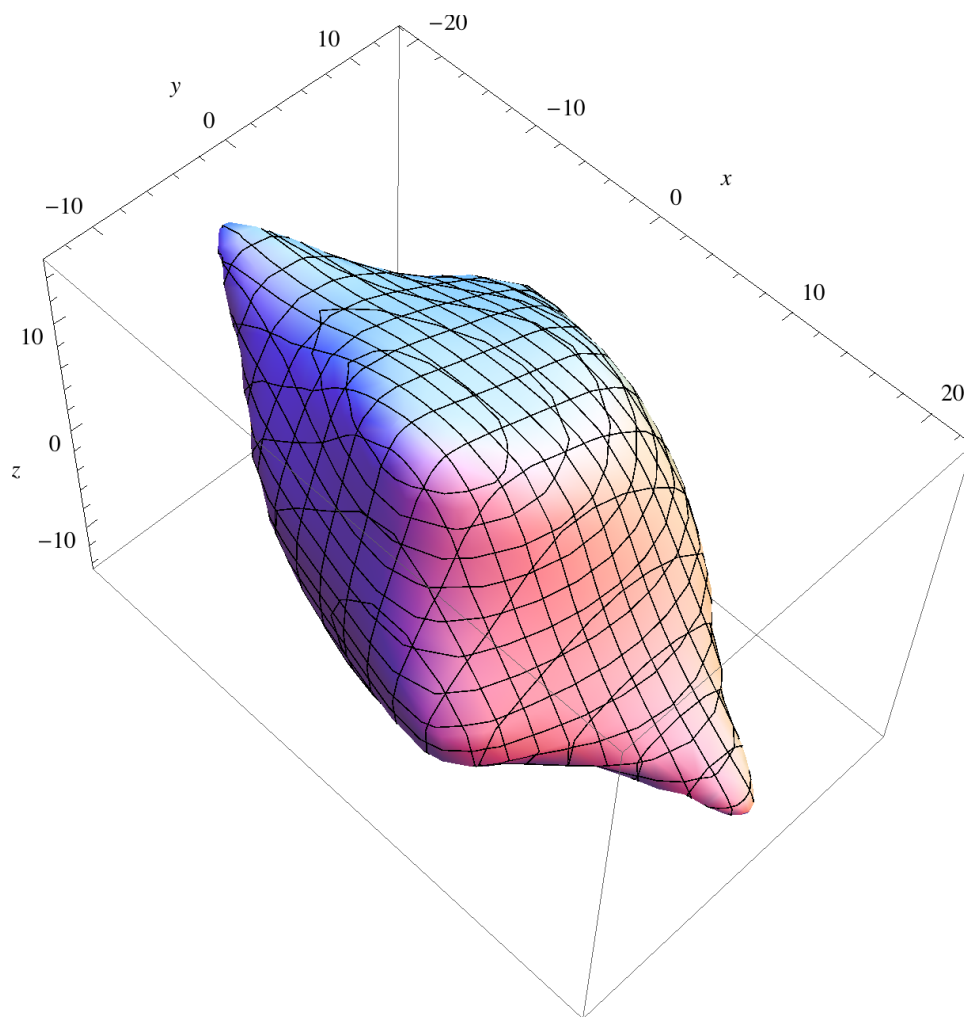


FIGURE 7. Contour Plot for the Surface of a Curvilinear Gaussian Potential.

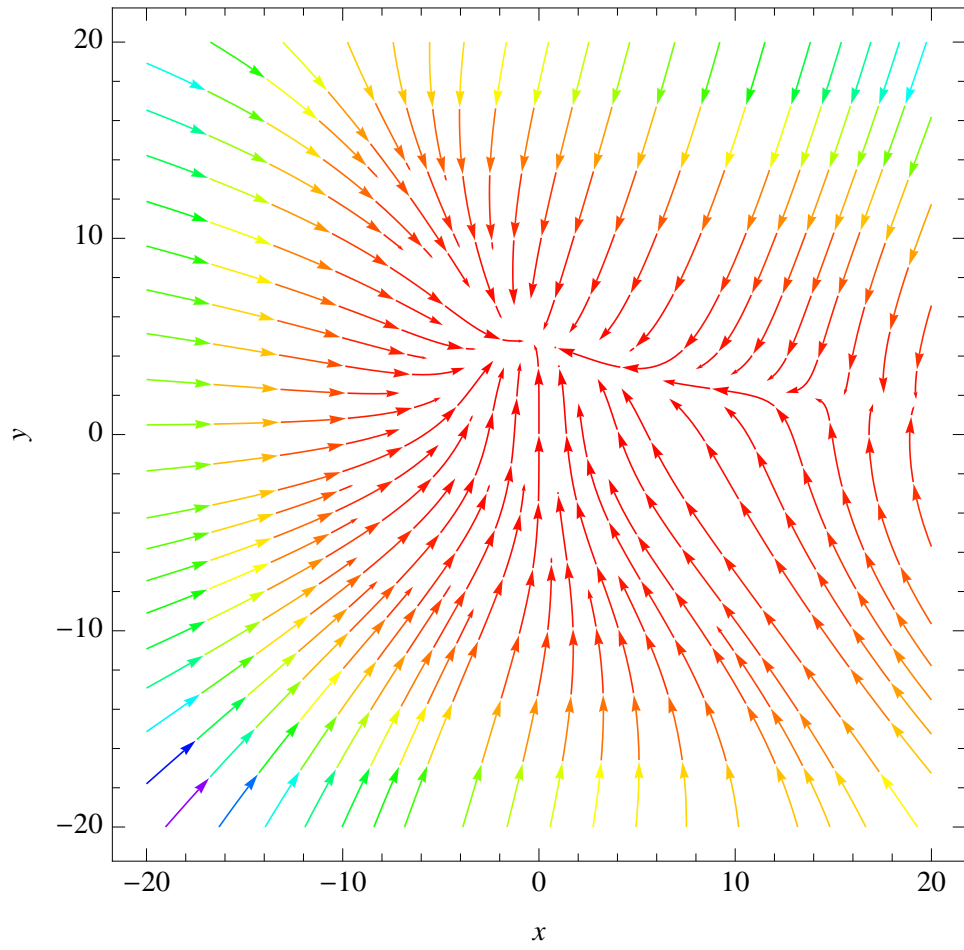


FIGURE 8. Gradient Vector Field for a Curvilinear Gaussian Potential at $z = -10$.

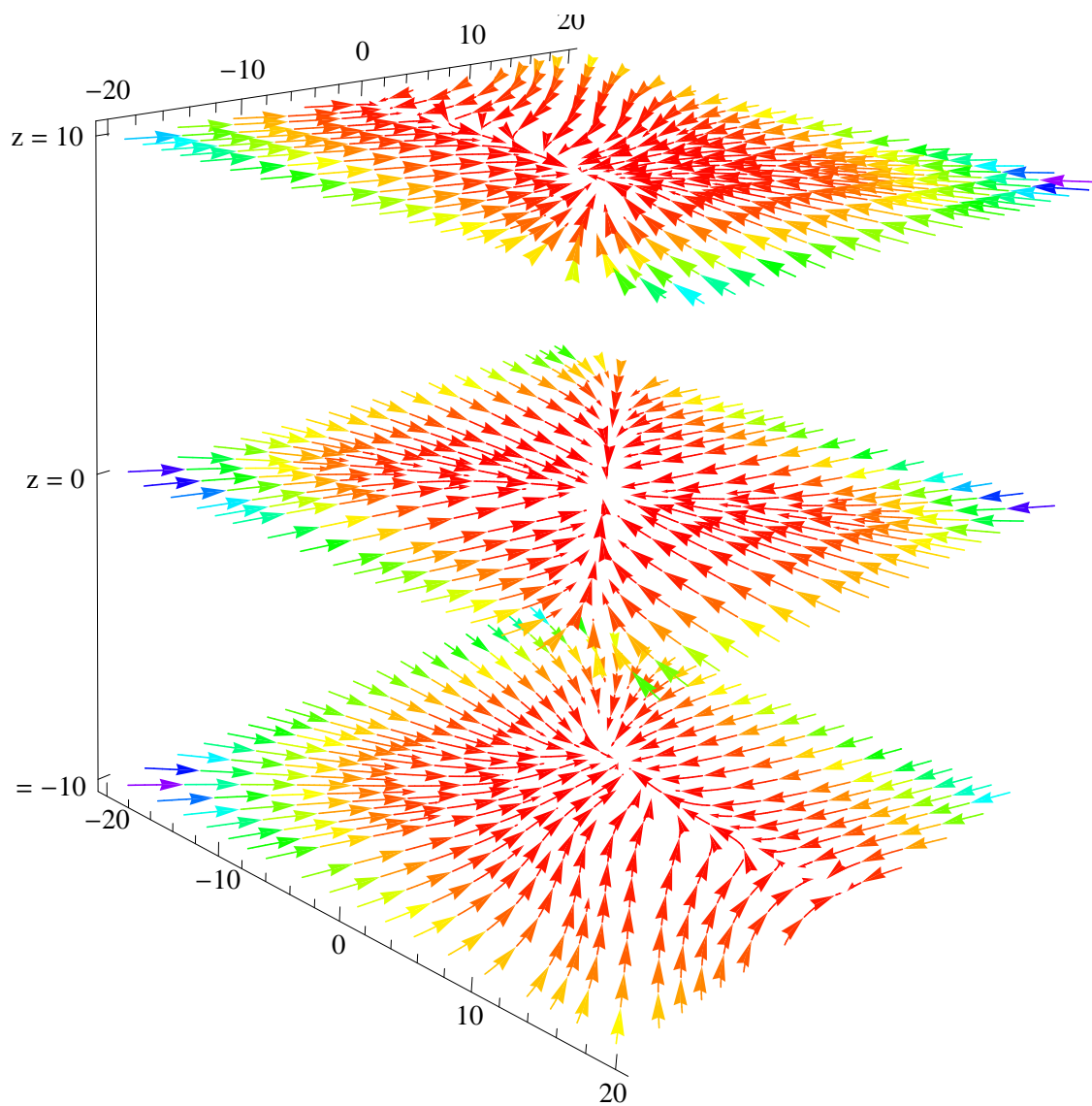


FIGURE 9. Gradient Vector Fields at $z = 10$, $z = 0$, $z = -10$.

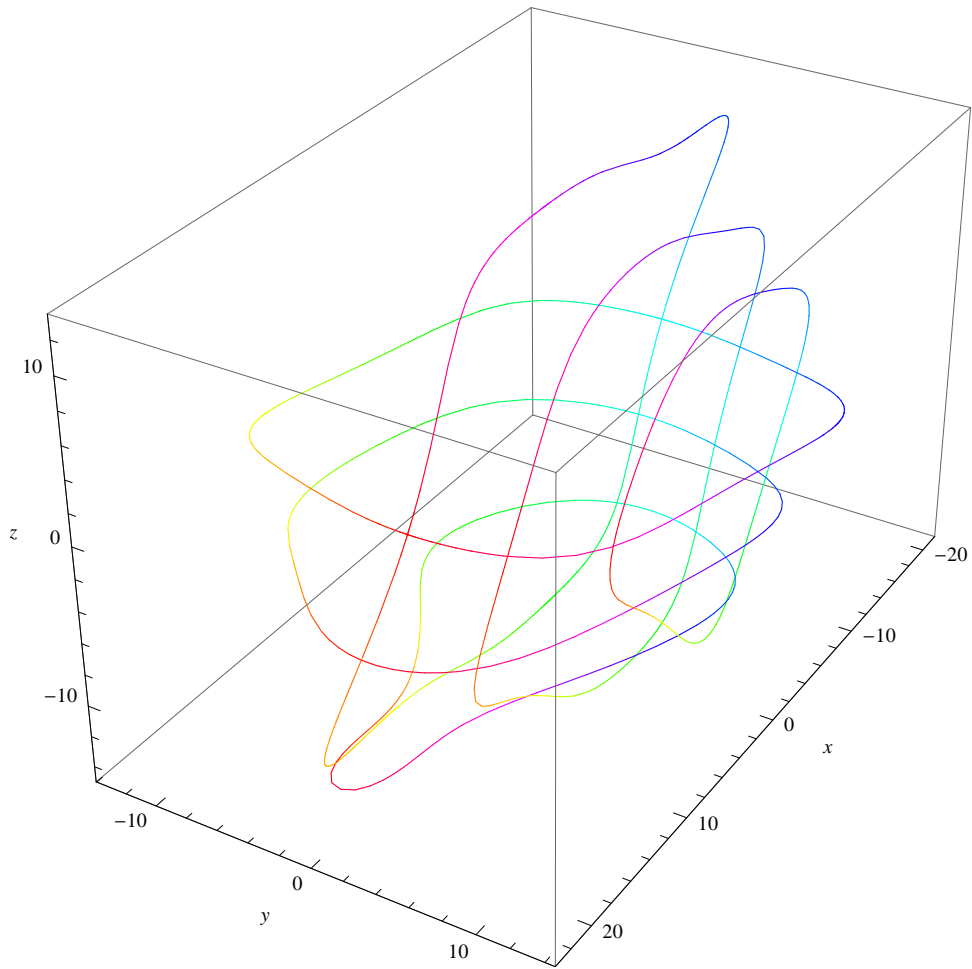


FIGURE 10. An Integral Manifold with a Curvilinear Coordinate System for a Curvilinear Gaussian Potential.

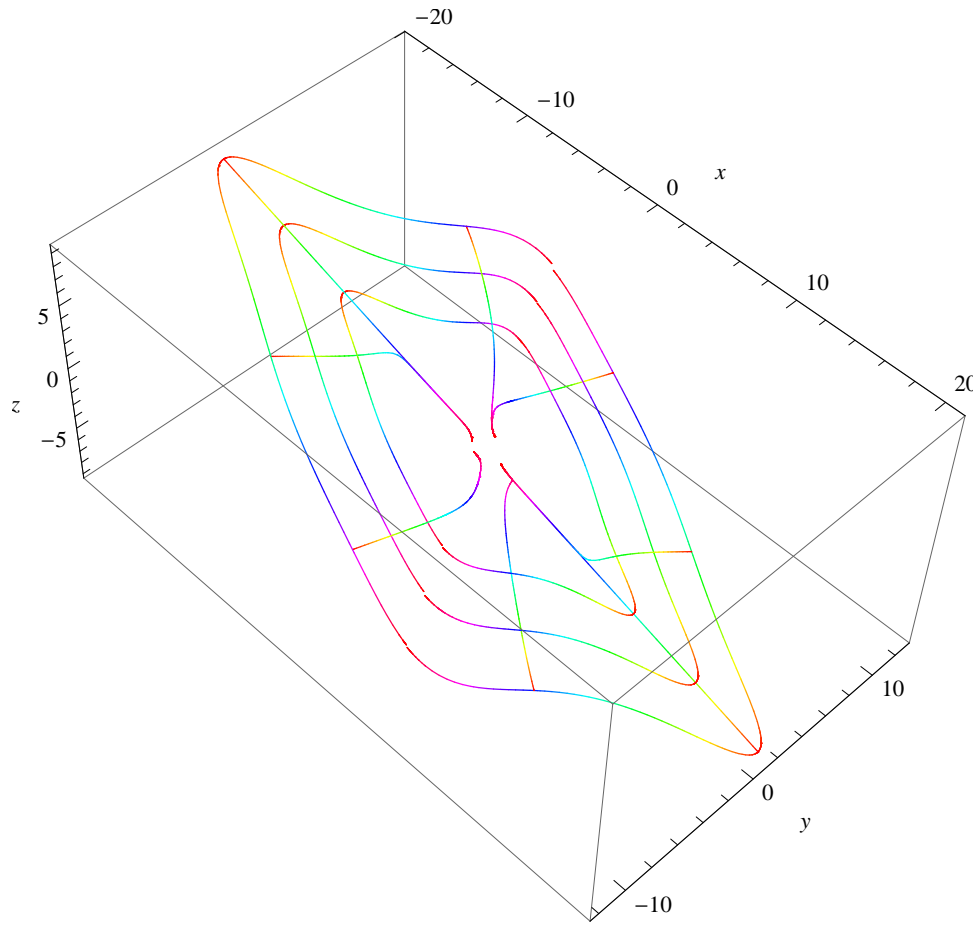


FIGURE 11. Curvilinear Coordinate System for the ρ, θ Surface of a Curvilinear Gaussian Potential, Using Two Antipodal Points.

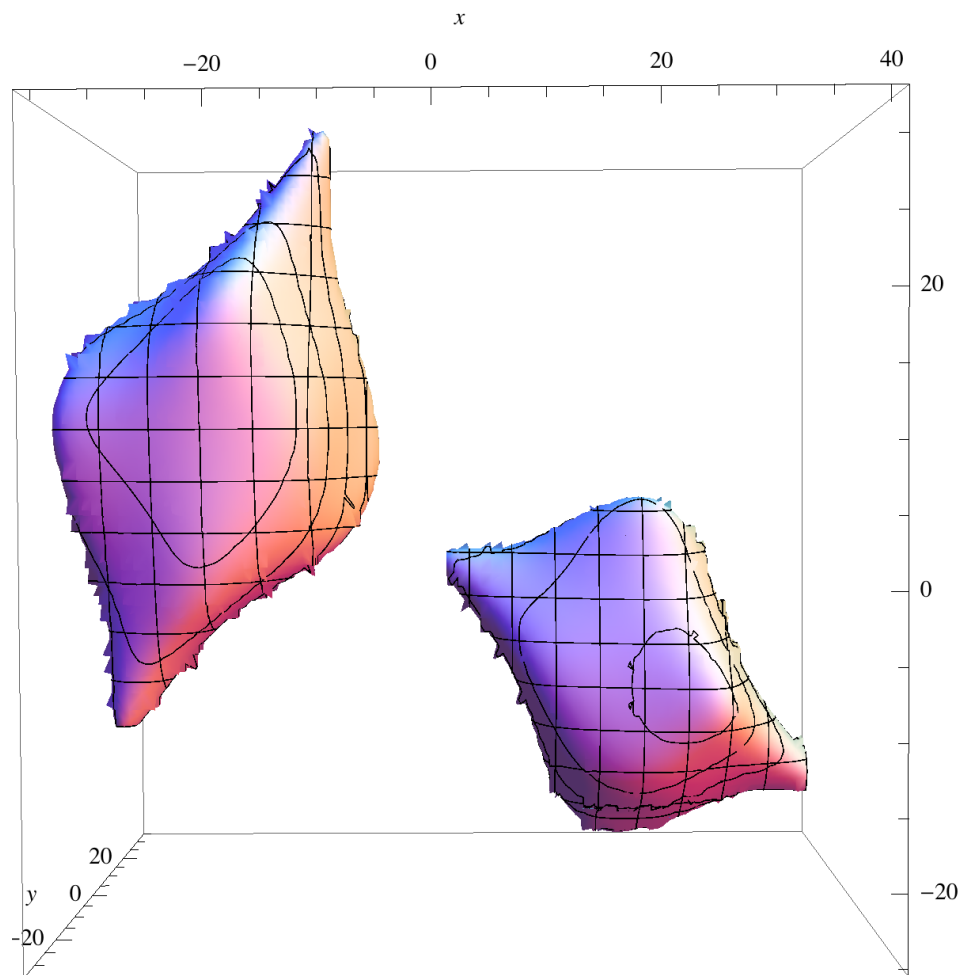


FIGURE 12. Mixture of Two Curvilinear Gaussians, Translated and Rotated.

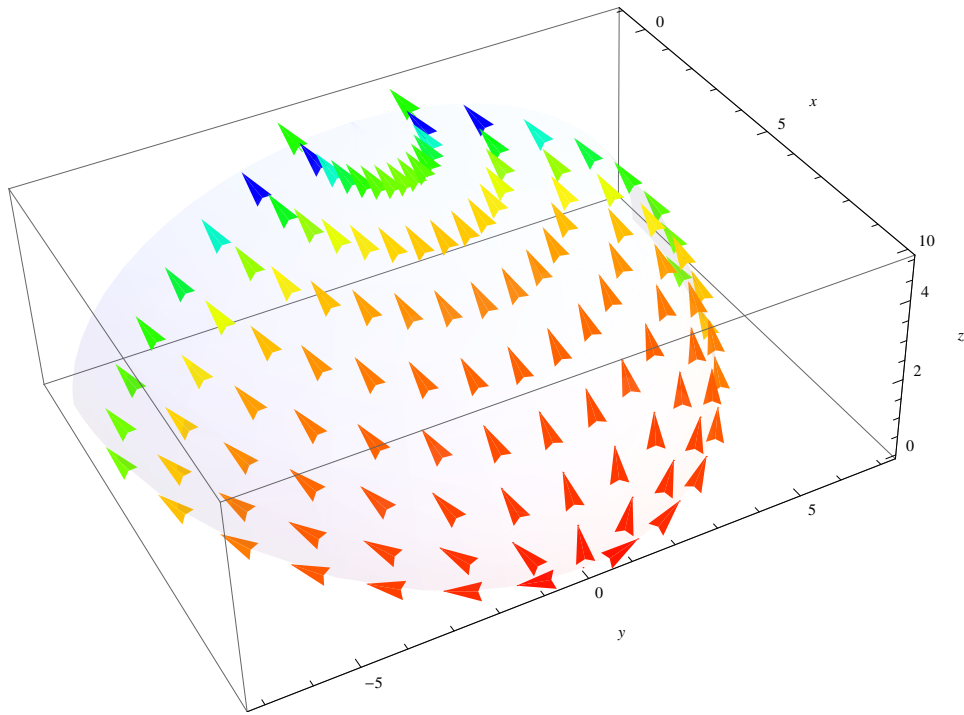


FIGURE 13. Drift Correction Vector Field for a Gaussian Diffusion.